Universidade de Aveiro Departamento de Matemática
2018

Mónica Joana
Morgado Rodrigues

# ANÁLISE DE RISCO NA ATIVIDADE FLORESTAL

## Risk analysis in forest activity

Mónica Joana
Morgado Rodrigues

# ANÁLISE DE RISCO NA ATIVIDADE FLORESTAL

## Risk analysis in forest activity

Relatório de Estágio apresentado à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, realizada sob a orientação científica de Maria da Conceição Cristo Santos Lopes Costa, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro e de Isabel Maria Simões Pereira, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro.

**o júri / the jury**

presidente / president

**Prof. Doutor Pedro Filipe Pessoa Macedo**
Professor Auxiliar do Departamento de Matemática da Universidade de Aveiro (por delegação da Reitora da Universidade de Aveiro)

vogais / examiners committee

**Prof. Doutora Maria Cristina Souto de Miranda**
Professora Adjunta do Instituto Superior de Contabilidade e Administração da Universidade de Aveiro (arguente)

**Prof. Doutora Maria da Conceição Cristo Santos Lopes Costa**
Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro (orientadora)

**agradecimentos / acknowledgements**

Reconhecendo o longo e árduo percurso na conclusão do mestrado, em especial na elaboração da presente dissertação, dedico este espaço àqueles que foram essenciais e infindáveis no apoio, direto ou indireto, a mim prestado. A todas essas pessoas aqui expresso os meus mais humildes e sinceros agradecimentos.

Ao RAIZ - Instituto de Investigação da Floresta e Papel, pela oportunidade proporcionada na realização do estágio curricular e desenvolvimento da investigação, a todas as pessoas com as quais convivi nesta instituição, em especial à Engenheira Margarida Silva, pelo acolhimento e apoio prestado ao longo de todo o período de estágio, tanto a nível institucional, tendo-se pautado sempre pelo rigor profissional, quanto ao nível pessoal, mostrando-se sempre disponível, agradecendo também ao Engenheiro João Gaspar, pela ajuda na compreensão dos conceitos e no processamento de dados, cruciais para o estudo desenvolvido nesta dissertação.

À orientadora desta dissertação, Professora Doutora Maria Conceição Lopes Costa, por todo o incentivo, orientação, disponibilidade, paciência, auxílio e compreensão ao longo de todas as atividades desenvolvidas.

À co-orientadora, Professora Doutora Isabel Pereira, pelos valiosos conhecimentos transmitidos, pelo elevado e rigoroso nível científico e pela visão crítica e construtiva, fulcrais na elaboração desta dissertação.

Aos meus pais, por todo o esforço que realizaram para me proporcionar a oportunidade de frequentar o ensino superior. A eles tudo devo.

À Catarina, amiga e companheira de estágio, por todo o apoio ao longo destes últimos anos, em especial nos tempos mais difíceis e trabalhosos com a elaboração desta dissertação.

Ao Raúl, por toda a paciência, apoio, carinho e sobretudo por estar sempre presente e disponível para me animar nos dias mais complicados.

Aos amigos, por fazerem parte da minha vida.

Por fim, mas não menos importante, a todas as pessoas que, de alguma forma, contribuíram para a escrita desta dissertação através de críticas, opiniões e sugestões. A todos o meu mais profundo agradecimento.

**Palavras-chave**

Análise de risco, autocorrelação, heteroscedasticidade, produtividade florestal, regressão robusta, variáveis *dummy*.

**Resumo**

Atualmente, a atividade florestal e a cadeia de produtividade a ela aliada assumem um importante papel na economia de Portugal, tornando-se crucial a formulação de estratégias e instrumentos que a apoiem. Considera-se neste trabalho o indicador de produtividade florestal Acréscimo Médio Anual em Volume, que suporta diversos processos de decisão em planeamento e gestão florestal tais como a idade de corte, a seleção do modelo de silvicultura e a exploração florestal. A modelação da produtividade florestal baseia-se em medições que refletem as condições médias que ocorreram no período de tempo em que as medições sucederam. As alterações climáticas e outros eventos, direta ou indiretamente relacionados com estas alterações, como o aumento da ocorrência de pragas e doenças ou do risco de incêndio traduzem-se em maior incerteza na obtenção de estimativas de produtividade e na tomada de decisão florestal.

Neste estudo pretende-se analisar de que forma o risco e a incerteza na ocorrência de uma das pragas que mais danos causa em povoamentos de eucalipto, o gorgulho do eucalipto, poderá afetar a estimativa da produtividade florestal em regiões de risco fraco a muito forte. No desenvolvimento da presente investigação, objetivando-se a implementação de um modelo que permita dar resposta ao problema colocado, foi feita uma análise de regressão linear múltipla, com inclusão de variáveis *dummy*. A análise do modelo construído permitiu detetar nos resíduos a presença de heteroscedasticidade e autocorrelação. Face à problemática referida, foi necessário aplicar métodos estatísticos adequados, entre os quais métodos de regressão robusta tais como regressão linear múltipla robusta e métodos baseados em estimadores consistentes na presença de heteroscedasticidade e autocorrelação.

**Keywords** Risk analysis, autocorrelation, heteroskedasticity, forest productivity, robust regression, *dummy* variables.

**Abstract**

Currently, forestry and the chain of productivity allied to it play an important role in the Portuguese economy, making it crucial to formulate strategies and instruments to support it. In this paper, it has been considered the Average Annual Average Volume forest productivity indicator that supports several decision-making processes in forestry planning and management, namely, age of cut, selection of forestry model and forest exploration. Forest productivity modeling is based on measurements that reflect the average conditions that occurred in a time period during which the measurements succeeded. Climate change and other events, directly or indirectly related to these changes, such as increased occurrence of pests and diseases or the risk of fire, translate into greater uncertainty in obtaining estimation of forest productivity and in decision-making.

The aim of this study is to analyze how the risk and uncertainty in the occurrence of one of the most damaging pests in eucalyptus, the eucalyptus weevil, could affect the estimation of forest productivity in regions with low to very high risk. In the development of the present investigation, aiming at the implementation of a model that allows answering the referred problem, a multiple linear regression analysis was done, including dummy variables. The analysis of the estimated model allowed detecting the presence of heteroskedasticity and autocorrelation in the residues. Considering the mentioned problem, it was necessary to apply suitable statistical methods, namely robust regression methods such as robust multiple linear regression and other methods based on consistent estimators in the presence of heteroskedasticity and autocorrelation.

# Conteúdo

# Lista de Figuras

# Lista de Tabelas

# Capítulo 1

# Introdução

A importância económica inerente à espécie *Eucapytus globulus* foi incrementada ao longo do século XX, devido à sua utilização como matéria-prima na indústria papeleira. Esta espécie foi introduzida em Portugal, por volta de 1829, com intuito de ornamentação, sendo que atualmente ocupa a maior área de floresta do território continental, o qual corresponde a cerca de 26% do território continental, isto é, 812 000 ha, (Associação da Indústria Papeleira – CELPA e Instituto Nacional de Investigação Agrária e Veterinária, 2015; Reboredo, 2014).

A rápida expansão inerente à espécie de eucalipto *globulus*, deveu-se fundamentalmente à ausência de pragas verificada ao longo dos anos, e, concomitantemente, ao seu uso na indústria papeleira, cujo crescimento foi substancial. As características inerentes a esta espécie de eucalipto, sobretudo a sua elevada taxa de crescimento e excelente qualidade da pasta de papel produzida, permitiu classificar esta espécie como sendo a espécie de eucalipto preferencial para plantações.

Com a afirmação de Portugal como um dos maiores produtores de pasta de papel da Europa, foi fundamental desenvolver metodologias de gestão das populações do eucalipto cada vez mais eficazes, capazes de maximizar a sua produtividade. Na deteção de pragas, mostrou-se ser essencial a implementação de técnicas de gestão florestal cada vez mais desenvolvidas, no sentido de minimizar os efeitos económicos provocados pelas pragas. Das várias espécies de insetos fitófagos que se alimentam de eucaliptos existentes no território nacional, o gorgulho do eucalipto tem vindo a ganhar destaque pelos danos nefastos a nível económico, não só em Portugal, mas à escala mundial. *Gonipterus platensis* é originário da Tasmânia sendo atualmente a espécie com a distribuição mais vasta, afetando regiões da Austrália, Europa, América do Norte, América do Sul e potencialmente África. Existe uma panóplia de espécies de *Eucalyptus* citadas como hospedeiras do *Gonipterus platensis*, especificamente, *E.viminalis*, *E.globulus* e *E.maindenii*, as quais são consideradas como preferidas pelo gorgulho.

Este inseto, tanto na fase larvar como na fase adulta, consome sobretudo as folhas, sendo que ataca principalmente o terço superior da planta hospedeira, causando uma diminuição da capacidade fotossintética da árvore e consequentemente do seu crescimento, o que por sua

vez acarreta consequências nefastas a nível económico. As larvas, após eclodirem, raspam a superfície foliar, alimentando-se da epiderme e mesófilo sem perfurar a epiderme oposta. Após o seu desenvolvimento, alimentam-se de forma indiscriminada de qualquer área das folhas e rebentações jovens. Os adultos tendem a alimentar-se preferencialmente nas extremidades das folhas e em ramos tenros, distribuindo-se de forma uniforme na planta, e podendo viver entre 6 a 12 meses.

Detetado pela primeira vez em Portugal, em 1995, na região do Norte, verificou-se que por volta de 2003 o *Gonipterus platensis* já se encontrava por todo o território português, com especial incidência nas regiões do Norte e Centro do país. A sua rápida propagação deveu-se, essencialmente, à ineficácia dos métodos de controlo implementados.

Com efeito, em 1995, aquando da sua descoberta, foi criado um projeto nacional de monitorização e controlo do *Gonipterus platensis*, tendo-se iniciado no ano seguinte as introduções das primeiras populações do parasitóide de ovos, *Anaphes nitens*, originário da Austrália. Analogamente ao que sucedeu noutros países, a sua introdução foi altamente eficaz em baixas altitudes e regiões com climas amenos. Contudo, em elevadas altitudes e regiões com invernos mais frios, esta medida não se mostrou profícua, devido ao desfasamento geográfico existente entre o parasitóide e o gorgulho, evidenciado por condições microclimáticas desfavoráveis. O facto de o parasitóide ser originário de climas mais amenos (Austrália do Sul) e o gorgulho de climas mais frios (Tasmânia) faz com que o primeiro tenha dificuldade em se adaptar a zonas ou alturas do ano com temperaturas mais baixas. Em Portugal, as zonas onde o dano é significativo correspondem às regiões Norte e Centro do país, que constituem cerca de um terço da área total de eucalipto. Tal fomentou a necessidade de encontrar novos métodos que complementassem a ação do parasitóide, sendo a alternativa o controlo químico. Contudo, a restrição no quadro legislativo em matéria de pesticidas, impede o uso da maioria dos produtos químicos em florestas e plantações certificadas, (Valente et al., 2004).

Uma opção ainda em estudo é a introdução de duas outras espécies de parasitóides provenientes da Tasmânia e que consequentemente apresentam uma resistência ao frio similar à do gorgulho. A produção de clones de *Eucalyptus globulus*, novas espécies ou híbridos menos suscetíveis ao ataque do inseto também tem vindo a ser desenvolvida, mas até agora não foi possível encontrar nenhuma alternativa pautada da mesma qualidade em termos de madeira e de floresta que o *Eucalyptus globulus*.

No âmbito da realização do estágio curricular conferente ao grau de Mestre em Matemática e Aplicações, no ramo de Estatística e Otimização, a empresa por mim escolhida foi o instituto de Investigação da Floresta e Papel, RAIZ. Os projetos desenvolvidos pelo referido instituto são maioritariamente da responsabilidade da multinacional The Navigator Company, cujo papel económico a nível mundial suscitou e motivou a minha escolha.

Assim, sendo a preocupação primordial da entidade de acolhimento o controlo e monitorização da praga e tendo por base as consequências nefastas que esta provoca nas suas plantações, o presente tema desenvolve um modelo capaz de prever a produtividade das plantações na presença da praga. Com efeito, o presente estudo prende-se com a necessidade de analisar de que forma o risco e a incerteza na ocorrência de uma das pragas que mais danos causa em povoamentos de eucalipto, o gorgulho do eucalipto, poderá afetar a estimativa da

produtividade florestal em regiões de risco fraco a muito forte. De facto, identificar corretamente a "doença" que atinge os eucaliptos e saber como os tratar de forma eficaz apresenta-se como primordial por forma a garantir uma plantação de eucaliptos saudável e com uma boa produção.

A presente investigação faz uso do Acréscimo Médio Anual em Volume (AMA, $m^3$/ha/ano) como indicador de produtividade florestal. Existe uma panóplia de fatores capazes de influenciar direta ou indiretamente a produtividade de uma parcela, a saber, alterações climáticas e outros eventos, nomeadamente a existência de pragas e doenças e o risco de incêndio. Estes eventos implicam uma maior incerteza na obtenção de estimativas de produtividade e na tomada de decisão florestal. Com efeito, o presente estudo pretende dar um contributo à necessidade de procura de um modelo capaz de estimar a produtividade de uma parcela, considerando apenas o risco inerente à presença da espécie invasora, *Gonipterus platensis*, (M. Branco et al., 2011; Reis et al., 2012).

A estrutura inerente à presente dissertação inicia-se pela apresentação da entidade de acolhimento, RAIZ, onde se evidencia o seu papel na economia nacional e mundial, enfatizando-se a sua importância ao nível da indústria papeleira.

Seguidamente, o capítulo 3, tem por objetivo tomar consciência das características inerentes à atuação do *Gonipterus platensis*, bem como, das consequências da existência da praga em Portugal Continental, particularmente em plantações do *Eucalyptus globulus*, visto que esta é considerada a espécie preferencial na produção de pasta e papel. Serão ainda analisadas metodologias de controlo da espécie invasora, tendo em conta que a referida problemática é o cerne da presente investigação.

No capítulo 4 são apresentadas as metodologias estatísticas que serviram de base à implementação de um modelo que permita dar resposta ao problema colocado. A sua apresentação prende-se com a necessidade de sustentar e consolidar as metodologias de inferência estatísticas aplicadas.

Por fim, no capítulo 5, foi feita uma análise de regressão linear múltipla, com inclusão de variáveis *dummy*. A análise do modelo construído permitiu detetar nos resíduos a presença de heteroscedasticidade e autocorrelação. Face à problemática referida, foi necessário aplicar métodos estatísticos adequados, que surgem como alternativa aos modelos de regressão clássicos, nomeadamente métodos de transformação das variáveis (transformação Box-Cox), métodos de estimação na presença de autocorrelação, métodos de regressão robusta tais como regressão linear múltipla robusta (RLMR) e métodos baseados em estimadores consistentes na presença de heteroscedasticidade e autocorrelação (*HAC - heteroskedasticity and autocorrelation consistent*).

A realização da presente investigação fez uso do *software* R (versão 3.4.2) e do *software* E.Views (versão 10 SV). Os dados foram processados usando o *software* QGIS.

# Capítulo 2

# Caraterização da instituição de acolhimento

O RAIZ - Instituto de Investigação da Floresta e Papel, formalmente criado em 1996 pela The Navigator Company, é uma entidade privada e sem fins lucrativos, cuja colaboração com instituições de ensino superior nacionais e estrangeiras, pela Europa, América e Austrália, permite o reforço da competitividade dos sectores florestal e papeleiro, por meio do apoio tecnológico e da formação especializada, com vista a desenvolver atividades de investigação, consultoria, serviços especializados e formação nos domínios da floresta, pasta, papel e biorrefinarias de base florestal. Esta instituição tem como principais sócios fundadores a Universidade de Coimbra, a Universidade de Aveiro, o Instituto Superior de Agronomia da Universidade de Lisboa e The Navigator Company, sendo este último o seu principal financiador, dentre outras entidades privadas, fundos públicos, tanto a nível nacional quanto europeu, sob uma perspetiva de produção e transformação dos conhecimentos técnico-científicos em produtos, serviços e tecnologia capazes de gerar mais-valias em termos competitivos para o setor silvo-industrial nacional, de modo a tornar a sua exploração mais eficiente e sustentável, (The Navigator Company, s.d.).

Neste contexto, atividade do RAIZ desenvolve-se com destaque para a Investigação Aplicada, sendo as áreas de intervenção florestal e tecnológica as que mais se salientam pelo desenvolvimento de parcerias estreitas com a indústria, em função de objetivos definidos, encetados sob a forma de projetos, conduzidos por gestores encarregues da coordenação de toda a atividade. As colaborações estabelecidas pelo RAIZ elevam e posicionam a instituição numa posição de prestígio, permitindo assim o sucesso das atividades desenvolvidas, sob uma visão científica de excelência, escrutinada pelo Conselho Científico existente na instituição. São também prestadas atividades no âmbito da consultoria florestal e tecnológica aos interessados, em geral, não sendo restrita às empresas associadas da instituição. Simultaneamente, a aposta na componente formativa dos seus trabalhadores é igualmente promovida na organização, com vista à criação de quadros altamente especializados em todos os ramos associados às áreas

das fileiras florestal e tecnológica, sobretudo ligadas ao estudo do eucalipto, salientando-se a vasta experiência no estudo de pragas, no melhoramento genético do eucalipto e biotecnologia, conferindo assim as capacidades técnicas e científicas adequadas para a execução dos projetos aos quais se propõe, (The Navigator Company, s.d.)..

Com efeito, o foco do RAIZ na obtenção de reconhecimento, a nível mundial, enquanto centro de investigação de excelência que procura fomentar o desenvolvimento sustentável e uma economia circular baseada na floresta de eucalipto, faz com que as atividades desenvolvidas neste âmbito tragam não só um incremento da produtividade florestal, como uma maior qualidade da fibra produzida, com vista a uma redução de custos de produção e uma menor pegada ambiental, sustentada através do melhoramento genético do eucalipto, assim como uma gestão florestal adaptada às diferentes regiões que comportam a produção de eucalipto, constituindo para tal o maior repositório mundial de conhecimento científico e tecnológico da espécie *Eucalyptus globulus*, predominante em Portugal, (The Navigator Company, s.d.). Para tal, a sede do RAIZ insere-se numa área com cerca de 2.600 $m^2$, na Quinta de S. Francisco, a 9 km de Aveiro, onde se desenvolve grande parte das atividades de investigação, dispondo de edifícios administrativos, laboratórios de bancada, instalações piloto, serviços centrais da sede, serviços de documentação, instalações para formação, instalações sociais. Possui ainda viveiros, parques de hibridação e um laboratório de biotecnologia, que se situam na Herdade de Espirra, em Pegões. Em termos de recursos humanos, responsáveis pelo desenvolvimento das atividades, a instituição dispõe de um quadro próprio de 52 investigadores e técnicos, para além de 25 bolseiros de I&D, em média, que se repartem entre os dois polos acima referidos (Quinta de São Francisco, em Eixo/Aveiro, e Herdade de Espirra, em Pegões), havendo uma articulação constante com as áreas operacionais (floresta e indústria) e corporativas da The Navigator Company, que incorporam uma rede alargada de universidades e centros de I&D, nacionais e estrangeiros, (Cluster Habitat Sustentável, 2018). O trabalho desenvolvido pela organização vê o seu êxito reconhecido, através de um galardão atribuído pela Ordem dos Engenheiros, bem como o seu reconhecimento como entidade do Sistema Científico e Tecnológico Nacional, distinção obtida pelos seus 20 anos de trabalho de investigação do eucalipto, (RAIZ, s.d.).

## 2.1 O grupo The Navigator Company

O grupo *The Navigator Company* detém uma participação de 94% no capital associativo do RAIZ, sendo o principal financiador das atividades desempenhadas pelo centro de investigação, dedicando-se ao fabrico e comercialização de papel em Portugal e no estrangeiro, de forma totalmente autónoma, no que diz respeito à produção da madeira, à pasta e ao papel, (The Navigator Company, 2016), (The Navigator Company, s.d.). Sendo responsável pela gestão de cerca de 3,2% da área florestal portuguesa e 14,4% da floresta nacional de eucalipto, o grupo é um dos principais responsáveis pela valorização da floresta do país, (Torrão, 2017), contando para tal com o maior viveiro de plantas florestais certificadas na Europa, cuja capacidade anual de produção ultrapassa os 12 milhões de plantas, sendo o RAIZ o centro de investigação responsável pelo envolvimento do grupo nos mais diversos projetos científicos com instituições públicas e privadas de referência a nível nacional e internacional.

O grupo constitui uma das mais fortes presenças na economia portuguesa, situando-se entre os três maiores exportadores de Portugal e um dos principais criadores de riqueza do país, representando cerca de 1% do Produto Interno Bruto nacional, exportando quase a totalidade da produção para 118 países dos cinco continentes. O sucesso do grupo deve-se, sobretudo, à forte aposta na Investigação e Desenvolvimento (I&D), o que desempenha um importante papel na manutenção e reforço da competitividade do seu modelo de negócio, baseado na investigação aplicada na floresta, pasta de celulose, energia renovável, papel e tissue, (Wikipédia, 2008). A gestão e o planeamento florestal são pilares fundamentais para a sustentabilidade do negócio da The Navigator Company, levando o grupo a encetar estratégias de implementação das melhores práticas neste âmbito, tendo sido reconhecido pelas certificações obtidas pelos sistemas internacionais FSC® (*Forest Stewardship Council®*) e PEFC™ (*Programme for the Endorsement of Forest Certification schemes*), demonstrando a importância atribuída à política de responsabilidade sustentável na prossecução dos objetivos do grupo, (The Navigator Company, 2016).

Deste modo, a atividade desempenhada pela The Navigator Company permitiu-lhe o reconhecimento como líder mundial no segmento *premium* de papéis de escritório com a marca *Navigator*, bem como produtor mundial de papéis finos de impressão e escrita não revestidos, e ainda de pasta branqueada de eucalipto, tendo sido distinguida, em 2013, pelo European Business Awards como a "Melhor Empresa da Europa"na categoria "Business of the Year", (The Navigator Company, 2016), (Grupo Portucel Soporcel, 2006). Os contributos do grupo vão para além da produção de pasta e papel, destacando-se igualmente na produção de energia, onde já representa mais de metade do total da energia produzida em Portugal a partir de biomassa, assim como 5% da produção total de energia elétrica produzida no país. (The Navigator Company, 2016)

# Capítulo 3

# Identificação da praga *Gonipterus platensis*

Com base nos dados do Inventário Florestal Nacional 2010 (IFN 2010), é plausível considerar o eucalipto *globulus* (*Eucaliptus globulus*) como sendo a espécie de eucalipto com maior presença em Portugal, ocupando cerca de 812 mil hectares. Esta espécie, originária da Austrália, e considerada espécie exótica em Portugal, representa o cerne e o suporte de sustentação da indústria de pasta e papel.

Apesar da sua importância reconhecida na indústria papeleira, a introdução do *Eucaliptus globulus* em Portugal acarreta vários riscos uma vez que é mais suscetível a pragas. Com efeito, no território nacional já foram identificadas 11 espécies de insetos australianos que se alimentam unicamente de eucalipto e que provocam estragos nas árvores da referida espécie.

Uma das pragas que fortemente ataca a espécie supracitada é o *Gonipterus platensis*, o gorgulho-do-eucalipto. Esta praga terá sido observada pela primeira vez em 1995 no norte de Portugal, sendo que apenas em 2003 terá sido confirmada a sua presença em todo o território nacional. Estudos demonstram que os maiores ataques têm sido observados no norte e centro do país, com particular incidência em eucaliptais instalados acima dos 500 metros de altitude, (Associação da Indústria Papeleira – CELPA e Instituto Nacional de Investigação Agrária e Veterinária, 2015).

Os estragos resultantes da presença deste parasita são identificados pela desfolha das árvores e pela redução do seu crescimento, podendo mesmo, em situações extremas de ataque, provocar a morte das árvores afetadas. Face ao panorama supracitado, em 2011 surgiu um plano de controlo cujo objetivo inerente consiste na minimização dos estragos e prejuízos causados pelos agentes bióticos nocivos. Nesse sentido, foram desenvolvidos dois meios de luta disponíveis: o meio de luta biológica, que utiliza os inimigos naturais; e a luta química que recorre a produtos fitofarmacêuticos no combate dessas pragas.

## 3.1 Caraterização

O *Gonipterus platensis* é uma praga desfolhadora oriunda da Austrália, que, em praticamente na totalidade do seu ciclo de vida se alimenta das folhas do eucalipto, cuja espécie preferencial é o *Eucalyptus globulus*. A praga recorre preferencialmente a folhas adultas recém-formadas. Por conseguinte, verifica-se que os eucaliptos mais suscetíveis ao ataque são os que se encontram em transição de folha jovem para adulta.

A germinação da praga depende do clima e do território existentes. Assim, no território nacional verifica-se que o *Gonipterus platensis* apresenta duas germinações por ano, a saber: na primavera e no outono. De facto, nas estações do ano referidas, verificam-se uma maior quantidade de posturas (ootecas – média de 8 ovos) e de larvas, relativamente às restantes fases do ciclo de vida do parasita. O seu ciclo de vida é bastante simples, podendo ser analisado a partir da figura 3.1.

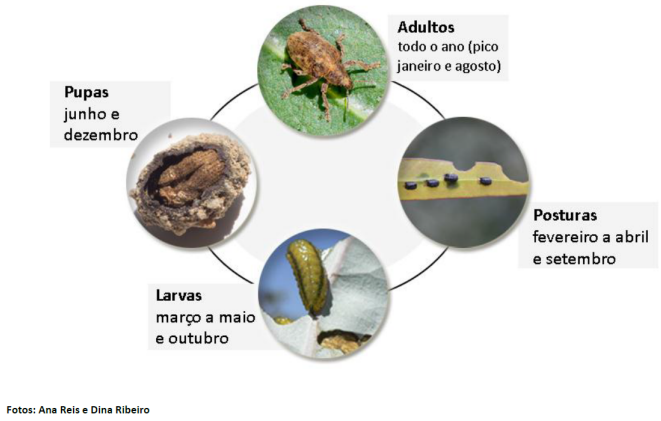



Figura 3.1: Ciclo de vida do *Gonipterus platensis*

A dispersão natural deste agente biótico nocivo pode ocorrer através do voo dos insetos adultos e é favorecida pelo movimento de material vegetal contaminado para plantação ou através do transporte de solo contaminado (larvas e pupas).

A nível mundial tem uma distribuição geográfica bastante alargada, tendo já sido detetado em diversos países (Mapondera et al., 2012), a saber: Espanha, Portugal, Ilhas Canárias, Estados Unidos da América (Califórnia) e Havai, Brasil, Argentina e Chile, Austrália e Nova Zelândia.

### 3.1.1 Sintomatologia e impacto económico

Os impactos provocados pelo parasitismo entre o *Gonipterus platensis* e o *Eucaliptus globulus* são diversificados, e de índoles variadas, traduzindo-se em consequências nefastas não só a nível económico e da produtividade da parcela, mas também a nível ambiental e paisagístico.

A deteção e observação dos sintomas pode realizar-se durante todo o ano, apesar de que existem épocas preferenciais, em particular, entre janeiro-abril e agosto-novembro. Os estragos provocados pela ação deste inseto apresentam maior visibilidade no terço superior da copa, onde surgem os novos rebentos. Desfolhas intensas e consecutivas reduzem fortemente o crescimento das árvores, o que se traduz numa perda total ou parcial da madeira utilizável. Em situações de ataque extremo, o ataque pode mesmo levar à morte das árvores. As consequências provocadas pelo parasitismo do *Gonipterus platensis* são bastante notórias, tendo originado prejuízos gravíssimos para a fileira do eucalipto e, por conseguinte, para a economia nacional.

Assim, não é recomendável permitir a propagação do ataque por áreas de estudo cada vez mais alargadas, e, por conseguinte, incrementar os efeitos nefastos que tal provoca em diversos âmbitos. Desta forma, torna-se imprescindível e urgente implementar o **Plano de controlo** para o inseto *Gonipterus platensis* (gorgulho-do-eucalipto) 2014-2015, de modo a minimizar, a curto prazo, o impacto desta praga, (Instituto da Conservação da Natureza e das Florestas, Direção-Geral de Alimentação e Veterinária, Instituto Nacional de Investigação Agrária e Veterinária, Associação da Indústria Papeleira – CELPA, Grupo Portucel Soporcel, Instituto de Investigação da Floresta, Papel – RAIZ e Altri florestal, 2015).

## 3.2 Monitorização

A decisão de implementar meios de controlo das populações de *Gonipterus platensis* depende, numa primeira fase, da intensidade do ataque (figura 3.2) e do vigor do arvoredo. A chave de decisão é apresentada na tabela 3.1.


Figura 3.2: Intensidade de ataque

Nas áreas previstas para tratamento com inseticida, deve ser feita a monitorização quinzenal do estado das árvores durante o período de maior atividade do inseto, com o objetivo de confirmar a necessidade de intervenção e determinar o momento mais adequado para a sua realização.

Tabela 3.1: Ação recomendada em função da intensidade do ataque (fonte: RAIZ).

| **Intensidade do ataque** | **Estragos observados** | **Ação recomendada** |
|---|---|---|
| Sem Ataque | Ausência de sinais da presença do inseto ou presença vestigial de estragos em poucas árvores. | Não intervir |
| Ataque Fraco | Presença de sinais vestigiais de alimentação na maior parte das árvores. As árvores mais atacadas têm menos de 20% de desfolha no terço apical. | Continuar a monitorizar |
| Ataque Moderado a Forte | Presença de estragos em todas as árvores, sob a forma de desfolha parcial, superior a 20% no ápice. Algumas árvores podem apresentar desfolha intensa, que pode chegar aos 90% no terço apical. Apesar da desfolha severa, as árvores mantêm a sua estrutura normal, i.e., copa cónica e tronco não deformado. | Aplicação de inseticida |
| Ataque Muito Forte | Desfolha muito intensa (>90%) ou total em todas as árvores. A maioria das árvores apresenta o tronco deformado, ramificado, com perda de dominância apical. Em povoamentos explorados em talhadia ocorre o aparecimento e desenvolvimento de varas secundárias e a perda de dominância por parte das varas selecionadas. É comum a existência de ramos secos ou de varas totalmente secas. | Corte raso e condução em talhadia ou replantação ou alteração do uso do solo |

## 3.3 Meios de luta

### 3.3.1 Luta biológica

Sendo o *Gonipterus platensis* uma espécie exótica, o controlo biológico clássico (introdução de inimigos naturais provenientes da sua região de origem), constitui uma estratégia de luta promissora. Embora exija um investimento relevante de recursos, depois dos inimigos naturais se estabelecerem com sucesso na área alvo, a limitação das populações da praga decorre naturalmente, sem necessidade de intervenção e de investimento adicional.

A luta biológica clássica tem sido a medida mais utilizada a nível mundial para controlo dos gorgulhos-do-eucalipto, mediante a introdução da espécie *Anaphes nitens* (*Hymenoptera: Mymaridae*) figura 3.3, um inseto com cerca de 1mm de comprimento que ataca os ovos de *Gonipterus platensis*.


Figura 3.3: *Anaphes nitens*

As fêmeas de *Anaphes nitens* depositam os seus ovos dentro dos ovos do gorgulho-do-eucalipto e as suas larvas destroem os embriões do gorgulho-do-eucalipto, o que, por conseguinte, afeta a taxa de reprodução da praga.

Apesar da espécie *Anaphes nitens* ser eficaz numa panóplia de regiões na redução das populações de *Gonipterus platensis*, existem locais cuja eficácia da espécie é reduzida, o que não evita a ocorrência de prejuízos nesses locais.

Assim, atendendo a que *Gonipterus platensis* e *Anaphes nitens* apresentam diferentes áreas de distribuição natural, coloca-se a hipótese de as duas espécies terem nichos climáticos diferenciados. Salienta-se ainda a utilização de inimigos naturais nativos da Tasmânia, potencialmente mais bem-adaptados às condições ambientais onde se verifica que *Gonipterus platensis* é praga, de forma a complementar a ação de *A. Nitens*. A estratégia supracitada poderá constituir uma estratégia de controlo eficaz.

Dos inimigos naturais de *Gonipterus platensis* até agora identificados na Tasmânia e testados em Portugal, o parasitóide *Anaphes inexpectatus* (*Hymenoptera: Mymaridae*) foi o único a ser multiplicado com sucesso em laboratório, usando *Gonipterus platensis* como hospedeiro,

sendo, pois, um bom candidato para estudo de eficácia.

Desde 2012, o Instituto RAIZ, a Altri Florestal e o Instituto Superior de Agronomia da Universidade de Lisboa desenvolvem, conjuntamente, trabalhos de investigação para avaliar a eficácia deste inimigo natural em campo. Entretanto, está a ser ponderado o pedido de autorização para importação e estudo de outros inimigos naturais de *Gonipterus platensis*.

### 3.3.2 Luta química

A aplicação de inseticida para controlo do gorgulho-do-eucalipto, deve ser realizada ao abrigo da Lei nº 26/2013, de 11 de abril, tendo em consideração os princípios da proteção integrada.

Só podem ser usados produtos fitofarmacêuticos autorizados para o fim em causa (cultura/inimigo), detendo para isso uma autorização de venda dada pela Autoridade Fitossanitária Nacional (DGAV). Devem ter-se ainda em consideração os produtos que estão proibidos pelos esquemas de certificação da gestão florestal e ser selecionados os que apresentem baixo risco, sobretudo para as abelhas.

Atualmente os produtos homologados para controlo do gorgulho-do-eucalipto são o *Calypso* e o *Epik*. Estes inseticidas são eficazes contra larvas e insetos adultos de gorgulho-do-eucalipto e não estão classificados como perigosos para as abelhas.

# Capítulo 4

# Enquadramento teórico

Na presente secção apresentam-se os fundamentos teóricos inerentes às análises estatísticas implementadas. Tal surge com o intuito de sustentar ao nível teórico e científico as metodologias aplicadas. Apresenta-se o teste de hipóteses de Kruskall Wallis utilizado nas comparações entre grupos. Seguidamente apresenta-se o modelo de regressão linear aplicado, bem como os testes de inferência estatística usados na avaliação dos resultados obtidos a partir da regressão.

Por fim, aborda-se a transformação de Box-Cox, métodos de estimação na presença de autocorrelação, a deteção de *outliers* e observações atípicas e, seguidamente, técnicas robustas de regressão linear, as quais surgem como alternativas ao procedimento de estimação OLS (método dos mínimos quadrados, do inglês *Ordinary Least Squares*), que visam contornar a violação dos pressupostos inerentes ao mesmo. Esta é uma área em ascensão e com um enorme relevo nas técnicas atuais de análise.

## 4.1 Teste não paramétrico de Kruskal Wallis

A aplicação de testes de hipóteses, independentemente de serem de índole paramétrica ou não paramétrica, justifica-se pela necessidade de analisar a existência de relações significativas no conjunto de dados.

O uso de testes não paramétricos justifica-se pela violação dos pressupostos inerentes aos testes paramétricos, em particular a falta de normalidade e a ausência de homogeneidade da variância dos dados.

O teste de Kruskal-Wallis (KW) é um teste não paramétrico usado para comparar no mínimo três ou mais populações. É comummente utilizado para testar a hipótese nula de que todas as populações têm funções de distribuição iguais contra a hipótese de que pelo menos duas populações têm função de distribuição diferentes.

Este teste é análogo ao teste F de Fisher, utilizado na ANOVA com um fator, porém, enquanto que na ANOVA a análise de variância dos testes depende da hipótese de que todas as populações em comparação são independentes e normalmente distribuídas, o teste KW não possui essas restrições.

A aplicação do método de Kruskal-Wallis, inicia-se com a ordenação das N observações do conjunto de dados, relativas às k amostras, por ordem crescente de observações. Suponha-se que $r_{ij}$ designa a posição de $X_{ij}$. Assim, toma-se

$$R_i = \sum_{j=1}^{n_i} r_{ij}, i = 1, ..., k, \tag{4.1}$$

onde $R_i$ é a soma das ordens de cada uma das $i$ $(i = 1, .., k)$ amostras. A estatística de Kruskal-Wallis H, será dada por

$$H = \frac{\frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)}{1 - \frac{\sum_{j=1}^{g}(t_j^3 - t_j)}{N^3 - N}}, \tag{4.2}$$

em que $t_j$ é o tamanho do grupo de elementos repetidos $j$ e $g$ é o número de grupos. Uma observação que não se repete é considerada como um grupo de tamanho 1. Esta estatística tem, aproximadamente, uma distribuição $\chi_{k-1}$, (Guo et al., 2013).

Sumariamente, o procedimento relativo à realização deste teste é:

1. Estabelecimento das hipóteses onde $\theta_i$ representa a mediana do grupo $i$, $i = 1, \ldots, k$,

$$\begin{aligned} &H_0 : \theta_1 = \theta_2 = ... = \theta_k; \\ &H_1 : \theta_1, \theta_2, ..., \theta_k, \text{ não são todas iguais.} \end{aligned} \tag{4.3}$$

2. Ordenação, de forma crescente de magnitude os valores deste novo conjunto de dados, e associação a cada valor a localização correspondente, sendo que cada posição possui o mesmo sinal do valor por este representado.

3. Cálculo do valor da estatística H. Em seguida, é fixado o nível de significância $\alpha$.

4. Obtenção dos valores críticos referentes ao nível de significância fixado. Neste caso, são calculados os valores $\chi_{\alpha-1}$ de modo que $P[H > \chi_{\alpha-1}] = \alpha$ (sob $H_0$).

5. Se $H_{obs} > \chi_{\alpha-1}$ rejeita-se a hipótese nula de que as amostras provêm de populações igualmente distribuídas.

6. O $p - value$ é calculado da seguinte forma $p - value = P[H \geq H_{obs}|H_0]$.

## 4.2 Regressão

O termo "Regressão" define um conjunto amplo de técnicas estatísticas empregues para modelar relações entre variáveis e predizer o valor de uma variável dependente (ou de resposta/explicada) a partir de um conjunto de variáveis independentes (ou preditoras/explicativas), (Marôco, 2010; Oliveira et al., 2011).

### 4.2.1 Regressão linear

No modelo de regressão linear a relação entre a variável dependente e as variáveis independentes pode ser descrita por uma reta, no caso de existir apenas uma variável independente, ou por um plano hipergeométrico, no caso de duas ou mais variáveis independentes. Quando existe apenas uma variável independente o modelo denomina-se modelo de regressão linear simples, caso contrário, designa-se regressão linear múltipla. Relativamente à variável dependente, quando a mesma é unidimensional denomina-se por regressão linear univariada, e caso contrário, o modelo diz-se de regressão linear multivariado, (Johnson, 81).

Visto que a investigação em curso tem como objetivo a análise da relação entre uma variável dependente e mais de que uma variável independente, o presente enquadramento teórico apenas incidirá sobre a regressão linear múltipla univariada.

Como já referido, o modelo de regressão linear relaciona uma variável dependente $Y$ com um conjunto de variáveis independentes $X_1, X_2, \cdots, X_k$, que influenciam a variável $Y$, e com uma variável aleatória $\varepsilon$, (Stapleton, 2009). Considerando que numa certa população, existe uma relação entre essas variáveis, é possível estabelecer a seguinte equação

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon, \tag{4.4}$$

em que $\beta_0, \beta_1, \beta_2, ..., \beta_k$ são constantes que representam os coeficientes de regressão que por sua vez representam os declives parciais e $\varepsilon$ designa o erro do modelo, que contabiliza erros de medição e efeitos de outras variáveis não explicitamente consideradas.

O modelo considerado anteriormente obriga a que os erros sejam aleatórios e não correlacionados, seguindo uma distribuição de probabilidades de média 0 e variância constante. Relativamente às variáveis independentes, o modelo exige que sejam ortogonais, por outras palavras, que não sejam correlacionadas e caso se verifique a existência de correlação que o valor seja próximo de zero. Os pressupostos supracitados serão explanados na secção 4.2.8.

A partir de um conjunto de dados, a construção de um modelo de regressão linear inicia-se pela estimação dos coeficientes de regressão. Para tal existem na literatura várias metodologias, nomeadamente o método dos mínimos quadrados, OLS, que se destaca pelas suas propriedades.

### 4.2.2 Regressão linear com variáveis categóricas

As variáveis utilizadas no modelo de regressão são, de um modo geral, quantitativas, porém é frequente encontrar fatores de índole qualitativa que influenciam o comportamento da variável dependente em estudo. A essas variáveis atribui-se a designação de variáveis categóricas, (Lewis-Beck, 1993).

O modelo de regressão contendo, como exemplo, uma variável dicotómica, pode ser representado na seguinte forma:

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i, \tag{4.5}$$

$$\text{com } D_i = \begin{cases} 1, & \text{se a observação de ordem } i \text{ se refere à categoria } c, \\ 0, & \text{no caso contrário.} \end{cases}$$

Note-se que o modelo escrito na equação 4.5 contém uma variável dicotómica que pode assumir apenas os valores 0 ou 1. Só assim é que o beta tem um índice fixo pois ou existe no modelo se $D_i$ for 1 ou não existe no modelo se $D_i$ for 0. Quando $D_i$ pode tomar outras categorias, o beta também tem de ter índice caso contrário teria que ter sempre o mesmo valor para todas as categorias.

Nesses casos, a presença ou ausência de alguma categoria, que se espera influenciar uma resposta, é considerada definindo uma variável que assume apenas os valores binários: o valor 0 faz com que o coeficiente desapareça da equação de regressão e o valor 1 faz com que o coeficiente atue como um termo constante adicional num modelo de regressão, representando a presença da categoria pretendida, (Oliveira et al., 2011). Deste modo, estas variáveis são usualmente denominadas por variáveis *dummy*, e este método de regressão é vulgarmente designado por regressão com variáveis *dummy*.

Em função da interação existente entre variáveis independentes usadas na regressão com variáveis *dummy*, é possível verificar a existência de três tipos de modelos: aditivo, multiplicativo e misto. Para facilidade de leitura vamos ilustrar estes três tipos de modelo no caso do modelo de regressão linear simples.

No modelo cuja variável categórica se encontra na forma aditiva, a equação de regressão com variáveis *dummy* é escrita na forma:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \varepsilon_i. \tag{4.6}$$

Em termos geométricos, é possível visualizar duas retas de regressão com o mesmo declive, mas ordenadas na origem diferentes, como se ilustra na figura 4.1 (pressupondo todas as estimativas dos coeficientes de regressão positivas). Esta situação considera, de forma implícita, que as diferenças nas categorias da variável categórica se mantêm constantes com a variação da variável quantitativa existente no modelo. Assume-se que esse diferencial se mantém exatamente o mesmo à medida que o valor da variável quantitativa aumenta/diminui, (Oliveira et al., 2011).

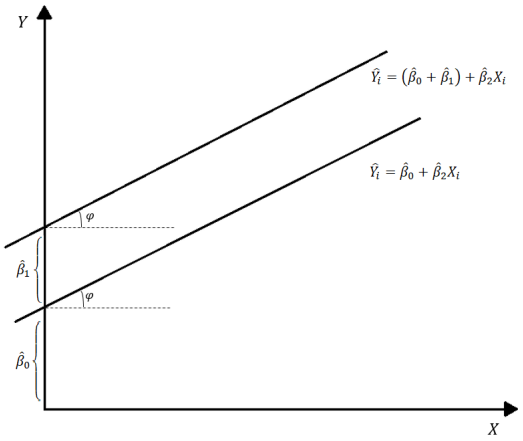



Figura 4.1: Modelo aditivo

Alternativamente, o modelo onde as variáveis categóricas se apresentam na forma multiplicativa, pode ser escrito na forma:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i X_i + \varepsilon_i. \tag{4.7}$$

Note-se que a introdução da variável binária na forma multiplicativa por $X_i$, tem o efeito de alterar o declive das duas retas estimadas, sendo que a ordenada na origem é a mesma. Com efeito, em termos geométricos as retas de regressão neste modelo podem ser visualizadas na figura 4.2. Assim, intuitivamente este modelo postula que a diferença entre as variáveis categóricas à medida que se verificam variações na variável quantitativa tende a variar.

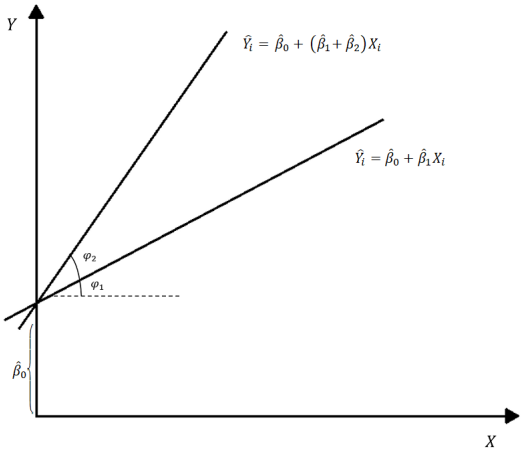



Figura 4.2: Modelo multiplicativo

O modelo de regressão com variáveis *dummy* na forma combinada aditiva e multiplicativa, permite garantir maior flexibilidade do modelo. A equação do modelo de regressão pode ser

escrita por:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 D_i X_i + \varepsilon_i. \tag{4.8}$$

Este modelo, em termos geométricos, possui ordenadas na origem e declives diferentes (figura 4.3). Assim, as ordenadas na origem permitem testar as diferenças significativas entre as categorias da variável *dummy* quando a variável quantitativa assume o valor nulo. Por outro lado, os declives das retas de regressão permitem inferir quanto às diferenças significativas entre as categorias da variável *dummy* com a evolução da variável numérica.

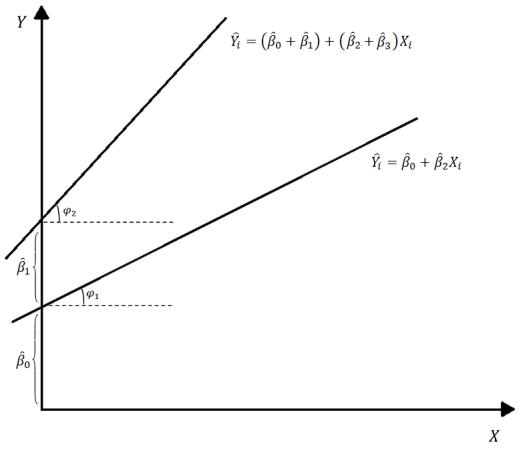



Figura 4.3: Modelo misto

### 4.2.3 Método dos mínimos quadrados - OLS

Como visto anteriormente, o primeiro objetivo na análise de regressão linear é o de estimar os coeficientes do modelo de regressão a partir de uma amostra representativa da população em estudo, usando os estimadores apropriados e tendo em conta a presença do termo de perturbação aleatório do modelo, (Lewis-Beck, 1993).

De acordo com o método dos mínimos quadrados, OLS, as estimativas dos coeficientes de regressão são obtidas de maneira a que a soma dos quadrados dos resíduos do modelo de regressão linear seja mínima, sendo os resíduos determinados por $e_i = Y_i - \widehat{Y}_i$.

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i} + ... + \widehat{\beta}_k X_{ki}, \tag{4.9}$$

em que $k$ é o número de regressores e $k+1$ o número de parâmetros.

O modelo, pode ser reescrito na forma matricial através de:

$$Y = X\beta + \varepsilon, \tag{4.10}$$

onde $Y$ representa o vetor das $n$ observações da variável dependente, $X$ é a matriz $(n \times (k+1))$ dos valores que as variáveis regressoras assumem, $\beta$ é o vetor dos coeficientes e $\varepsilon$ é o vetor dos erros ou termos de perturbação aleatória.

Como o próprio nome indica, o método tem como objetivo selecionar o estimador que minimiza a soma dos quadrados das distâncias entre os valores observados da variável resposta e aqueles que são preditos pelo modelo, (Brown, 1993), por outras palavras, minimiza a soma dos quadrados dos resíduos

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2, \tag{4.11}$$

onde $n$ representa o número de observações da amostra.

Admitindo que a matriz $X^T X$ tem determinante não nulo, o estimador de $\beta$ conseguido pelo método referido é da forma

$$\widehat{\beta} = (X^T X)^{-1} X^T Y, \tag{4.12}$$

onde $X^T$ representa a matriz transposta de $X$.

### 4.2.4 Método de seleção de preditores

Num problema de regressão linear, à partida conhecem-se quais as variáveis independentes a incluir no modelo. Todavia, particularmente em fases exploratórias da regressão linear, pode desconhecer-se qual ou quais as variáveis que conduzem ao melhor modelo, e, por conseguinte, a decisão da seleção de variáveis pode ser complicada pela presença de colinearidade e dos seus efeitos sobre a magnitude e o sinal dos coeficientes da regressão, (Johnson, 81; Marôco, 2010; Yamashita et al., 2007).

Existem vários métodos para a seleção do melhor modelo, sendo eles:

- **Seleção *Forward*:** principia com o modelo sem variáveis independentes e, a cada passo, introduz a variável que mais favorece o modelo no que se refere ao valor da estatística F. Se esta for inferior a um determinado valor estabelecido essa variável é eliminada e procede-se à avaliação da próxima variável; no final escolhe o melhor modelo entre a totalidade dos modelos construídos;

- **Seleção *Backward*:** inicia com o modelo com todas as variáveis independentes e, a cada passo, retira a variável por forma a que essa eliminação permita melhorar o valor da estatística F, e, por conseguinte, melhorar o modelo; no final escolhe o melhor modelo de entre todos os produzidos;

- **Seleção *Stepwise*:** este procedimento é um híbrido dos dois explicados anteriormente. No primeiro passo a seleção *Stepwise* inicia-se só com uma variável independente (como no método *Forward*) contudo a significância de cada adição de uma nova variável independente ao modelo é testada como no método *Backward*. A vantagem deste método, é

que permite a eliminação de uma variável cuja importância no modelo é reduzida pelo incremento de novas variáveis. Este procedimento acaba quando nenhuma das variáveis independentes ainda de fora, consegue entrar no modelo, e nenhuma das variáveis independentes presentes no modelo é expulsa, com base no critério de comparação de modelos.

### 4.2.5 Avaliação de significância de uma variável explicativa

Um dos testes mais relevantes na análise estatística do modelo de regressão linear é o teste de significância de uma variável explicativa. Consiste em verificar se algum dos coeficientes das variáveis explicativas é igual a zero. Se tal acontecer, a variável explicativa referente ao coeficiente que toma o valor zero não será explicativa da variável dependente, (Oliveira et al., 2011).

De forma resumida, testar o efeito da variável explicativa $X_j$ sobre a variável explicada $Y$, equivale às seguintes hipóteses:

$$H_0 : \beta_j = 0 \quad vs \quad H_1 : \beta_j \neq 0, \quad j = 1, 2, ..., k.$$

Assumindo que os erros seguem distribuição $N(0, \sigma^2)$, se $\sigma^2$ for desconhecida e $H_0$ verdadeira, a estatística de teste designada por estatística t, é determinada através de

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{(n-k-1)}. \tag{4.13}$$

Para um dado nível de significância $\alpha$, rejeita-se $H_0$ se $|t_{obs}| > t_{1-\frac{\alpha}{2},n-k-1}$ e conclui-se pela significância individual da variável explicativa $X_j$. Por outro lado, quando não se rejeita $H_0$, isto é, $\beta_j = 0$ é possível afirmar que para um determinado nível de significância e pela estatística t, que a variável $X_j$ não é estatisticamente significativa.

### 4.2.6 Avaliação global da significância do modelo

Tendo em conta o modelo de regressão linear da equação 4.9, é possível verificar que, se todos os coeficientes associados às variáveis explicativas do modelo forem simultaneamente iguais a zero, as variáveis explicativas serão, no seu conjunto, consideradas estatisticamente não significativas para explicar a variável dependente, $Y$, (Oliveira et al., 2011).

Desta forma, para testar a significância global do modelo de regressão surge o teste F, que consiste na comparação da soma dos quadrados dos erros do modelo de regressão múltipla com a soma de quadrados dos erros do modelo de regressão no qual a hipótese nula é admitida como verdadeira, para as hipóteses:

$$H_0: \ \beta_1 = \beta_2 = ... = \beta_k = 0 \quad vs \quad H_1: \exists \ \beta_j \neq 0, \quad j = 1, 2, ..., k.$$

A estatística F é dada por:

$$F = \frac{SQE/(k)}{SQR/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)} \sim F_{(k,n-k-1)}, \tag{4.14}$$

onde $SQE = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ é a soma de quadrados explicada e $SQR = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ é a soma de quadrados dos resíduos de regressão. A estatística apresentada segue a distribuição F de Fisher, com $k$ graus de liberdade no numerador e $n-k-1$ graus de liberdade no denominador, onde $n$ e $k$ representam, o número de observações do conjunto de dados e o número de variáveis explicativas usadas no modelo de regressão (isto é, regressores), respetivamente, assumindo que os erros têm distribuição Normal.

Considerando um dado nível de significância, rejeita-se $H_0$ se $F_{obs} > F_{1-\alpha,(k,k-n-1)}$, e nesse caso é possível afirmar que as variáveis explicativas se relacionam com a variável independente de uma forma estatisticamente significativa, por outras palavras afirma-se a significância global do modelo. Caso contrário, se $H_0$ não é rejeitada, constata-se que a regressão não é globalmente significativa para explicar a variável dependente, para o determinado nível de significância.

Atente-se que, relativamente a um conjunto de dados de elevada dimensão, valores da estatística F mesmo que um pouco distanciados de 1 podem levar a uma rejeição extemporânea da hipótese nula, o que leva a comprovar a relação entre as variáveis independentes e a variável de resposta, (Hair et al., 2010; Hamburg & Young, 1994).

Ao contrário, para conjuntos de dimensão reduzida, para que exista uma relação entre as variáveis explicativas e a variável dependente, os valores da estatística F precisam de ser elevados.

Este teste somente é válido para modelos com termo independente.

### 4.2.7 Critério de comparação de modelos – $R^2$, AIC e BIC

Os critérios comummente usados na seleção de variáveis incluem o cálculo do coeficiente de determinação múltiplo ($R^2$), o Critério de Informação de Akaike (AIC – *Akaike Information Criterion*) e o Critério de Informação Bayesiano (BIC – *Bayesian Information Criterion*).

#### Coeficiente de determinação múltiplo – $R^2$

Coeficiente de determinação múltiplo – $R^2$, representa a proporção de variação da variável dependente que é explicada pela(s) variável(eis) independente(s), em outras palavras, é uma

medida alusiva ao poder explicativo do modelo usado e é calculado por

$$R^2 = 1 - \frac{SQR}{SQT}, \tag{4.15}$$

onde $SQT = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ é a soma de quadrados total.

A avaliação do modelo por este critério consiste na interpretação do valor resultante, que está compreendido entre 0 e 1: quanto mais próximo for de um, maior é a quantidade de variação da variável dependente que é explicada pela(s) variável(eis) independente(s), melhor dizendo, maior poder explicativo possui o modelo. Caso o valor de $R^2$ seja próximo de zero, suscita à inadequação do modelo utilizado.

Uma caraterística importante deste critério é que este tende a aumentar na medida em que se incrementa uma nova variável independente, mesmo que a influência exercida sobre a variável dependente seja diminuta. Posto isto, a utilização do coeficiente de $R^2$ não é recomendada para a comparação de modelos com diferente número de variáveis independentes.

Por forma a colmatar este problema, surge o coeficiente de determinação ajustado, $R^2_{ajust}$, tratando-se de um critério mais adequado uma vez que este apenas aumenta se o acréscimo da nova variável produzir um melhor ajustamento do modelo aos dados. Este coeficiente é dado por

$$R^2_{ajust} = 1 - \frac{SQR/(n-k-1)}{SQT/(n-1)}, \tag{4.16}$$

onde $n$ e $k$ representam o número de observações da amostra e o número de regressores do modelo, respetivamente. Salienta-se portanto, que ao contrário do coeficiente $R^2$ este coeficiente é capaz de lidar com a adição de uma nova variável, ou seja, aumenta quando a variabilidade dos erros diminui relativamente à variabilidade total. (Murteira et al., 2010)

### Critério de informação de Akaike – AIC

Segundo Yamashita et al. o valor do critério de informação de Akaike – AIC, denota vantagens comparativamente a outros critérios comummente usados, nomeadamente ao anterior critério, uma vez que não não se baseia na decomposição da variabilidade. Este valor é obtido para cada modelo do conjunto de modelos possíveis, através da seguinte equação:

$$AIC = 2(k+1) - 2\ln L, \tag{4.17}$$

onde $k+1$ denota o número de parâmetros do modelo e $L$ o máximo valor da função de verosimilhança da amostra para o modelo estimado. O modelo mais adequado mediante a aplicação deste critério é o que apresenta um menor valor de AIC, (Guimarães & Cabral, 2010).

#### Critério de informação Bayesiano – BIC

Por fim, salienta-se o critério BIC na comparação de modelos, o qual se baseia na função de verosimilhança, encontrando-se intimamente relacionado com o AIC. Ambos os critérios utilizam um termo de penalização para o número de parâmetros do modelo, sendo que o termo de penalização é maior no BIC que no AIC. O valor do BIC é calculado através da seguinte equação:

$$BIC = (k+1)\ln n - 2\ln L, \tag{4.18}$$

onde $n$ representa o número de observações do modelo e os restantes parâmetros estão de acordo com o explicado anteriormente.

### 4.2.8 Validação dos pressupostos do modelo de regressão linear

O modelo de regressão linear, exposto anteriormente, só pode ser usado com objetivos de estimação e de inferência de relações funcionais entre a variável dependente e as variáveis independentes, se um conjunto de pressupostos respeitantes ao modelo forem válidos. Assim, a regressão linear prossegue, comummente após a estimação dos coeficientes da regressão, com a validação dos pressupostos respeitantes aos erros ou termos de perturbação aleatória e à (aproximada) ortogonalidade entre as variáveis independentes, (Oliveira et al., 2011).

Note-se que os erros não são conhecidos e por isso a validação dos pressupostos é feita sobre os resíduos que se aceitam como estimativas para os erros.

#### Análise de resíduos

No modelo de regressão linear, os resíduos ($e_i$) do modelo serviam quer para estimar os coeficientes de regressão quer para validar os pressupostos de aplicação do modelo de regressão linear. Efetivamente, a inferência acerca do modelo só é válida quando se verifiquem as condições explanadas seguidamente, (Marôco, 2010).

As hipóteses clássicas que caraterizam os principais momentos da distribuição do termo de perturbações aleatórias, $\varepsilon$, são:

$H_1$ : O valor esperado do termo de perturbações é 0. Traduz-se na seguinte equação:

$$E[\varepsilon_i] = 0, \quad \forall i. \tag{4.19}$$

$H_2$ : A variância do termo de perturbação é constante.

Isto refere-se à hipótese de homocedasticidade (igual variância), que pode ser escrita como:

$$Var[\varepsilon_i] = \sigma^2, \quad \forall i, \quad 0 < \sigma^2 < \infty. \tag{4.20}$$

$H_3$ : Duas quaisquer perturbações aleatórias não estão correlacionadas.

Esta hipótese diz respeito à hipótese de ausência de autocorrelação, que pode ser expressa na equação:

$$Cov(\varepsilon_i, \varepsilon_j) = 0, \quad \forall i, j : i \neq j. \tag{4.21}$$

Nos pontos seguintes serão explanadas com mais detalhe as hipóteses mencionadas anteriormente e quais as consequências das suas violações.

### 1. A hipótese da normalidade

Como vimos anteriormente, as hipóteses clássicas $[H_1]$, $[H_2]$ e $[H_3]$ caraterizam os principais momentos da distribuição de $\varepsilon$ sem depender do tipo de distribuição de probabilidade envolvida. O mais comum é assumir que o termo de perturbações $\varepsilon$ segue distribuição Normal n-variada:

$$[H_4] : \varepsilon \sim N_n(0, \sigma^2 I_n). \tag{4.22}$$

Uma conjunção de $[H_4]$ com $[H_1]$, $[H_2]$ e $[H_3]$ permite concluir que os $\varepsilon_i \quad i = 1, 2, ..., n$ são variáveis aleatórias independentes e identicamente distribuídas com distribuição Normal de média 0 e variância $\sigma^2$, (Mardia et al., 1994),

$$\varepsilon_i \sim N(0, \sigma^2), \tag{4.23}$$

conduzindo às seguintes consequências:

(i) A variável dependente Y segue distribuição Normal n variada, cujas marginais $Y_i$ têm distribuição também Normal dada por:

$$Y_i \sim N(\beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki}, \sigma^2) \quad \forall i = 1, 2, ..., n. \tag{4.24}$$

(ii) O estimador OLS de $\beta$ segue distribuição Normal de componentes:

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 a_{ij}) \quad \forall j = 1, 2, ..., k+1, \tag{4.25}$$

sendo $a_{ij}$ o elemento da linha $i$ e da coluna $j$ da matriz simétrica $(X^T X)^{-1}$, referida em 4.12.

(iii) Os resíduos de estimação $e_i$ seguem distribuição Normal:

$$e_i \sim N(0, \sigma^2 m_{ii}) \quad \forall j = 1, 2, ..., n, \tag{4.26}$$

sendo $m_{ii}$ o elemento da linha $i$ e da coluna $i$ da matriz simétrica
$M = I_n - X(X^T X)^{-1} X^T$.

(iv) A variável aleatória $\frac{(n-k-1)\hat{\sigma}^2}{\sigma^2}$ segue distribuição Qui-quadrado com $n-k-1$ graus de liberdade.

$$\frac{(n-k-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k-1}, \tag{4.27}$$

com $\hat{\sigma}^2 = \frac{e^T e}{n-k-1}$.

(v) Sob a equação 4.10 e a hipótese $[H_4]$ os estimadores OLS são estimadores BUE (*Best Unbiased Estimators*), isto é, os estimadores OLS são os estimadores de variância mínima dentro da classe dos estimadores cêntricos.

Aquando da violação da hipótese de normalidade, os estimadores OLS continuam a ser centrados e consistentes, só que deixam de ser os mais eficientes (os de variância mínima) e a inferência estatística deixa de ser válida (em particular, deixam de ser válidos os intervalos de confiança e os testes de hipóteses sobre os coeficientes individuais, que estão relacionados com a significância estatística das variáveis explicativas), (Stapleton, 2009).

A normalidade dos resíduos pode ser verificada através de métodos gráficos, através de um *QQ-plot* (gráfico de quantis Normal) dos resíduos, em que as observações se devem aproximar da bissetriz dos quadrantes ímpares (com dados não estandardizados), ou usando testes de normalidade.

Os testes formais de normalidade enunciados são testes que comparam a função de distribuição empírica que é estimada com base nos dados, com a função de distribuição cumulativa da distribuição normal. Estes testes designam-se por testes EDF (*Empirical Distribution Function*). Dufour et al. descreveu os testes EDF como testes baseados numa medida de desfasamento entre as distribuições empíricas e hipotéticas, (Razali & Wah, 2011). Os testes mais comummente usados são:

- **Teste de *Shapiro-Wilk*:** Um dos primeiros testes apto para diagnosticar a normalidade de uma amostra, foi proposto por Shapiro & Wilk em 1965. Este teste salienta-se pela obtenção de bons resultados, nomeadamente em amostras de dimensão inferior ou igual a 50 observações, (Leotti et al., 2012).

  O teste de Shapiro-Wilk tem como propósito aferir se os dados de uma amostra são ou não provenientes de uma distribuição Normal. Dada uma amostra aleatória ordenada $X_1 < X_2 < ... < X_n$, a estatística de teste inerente é:

  $$W = \frac{\left(\sum_{i=1}^{n} a_i X_i\right)^2}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2},$$

  onde $X_i$ designa a $i$-ésima observação da variável $X$ na amostra; $\bar{X}$ é a média da variável $X$ na amostra e $a_i$ são as constantes geradas a partir da média, variância e covariância de ordem $n$ com a distribuição Normal.

  Para diferentes níveis de significância $\alpha$ e diferentes tamanhos da amostra $n$, os valores críticos desta estatística de teste ($W_{\alpha,n}$) encontram-se tabelados, com valores compreendidos entre 0 e 1. Para a rejeição de $H_0$ a um nível de significância $\alpha$ tem-se $W_{obs} \leq W_{\alpha}$, (Rahman & Govindarajulu, 1997).

  Sendo este teste apenas possível para amostras de dimensão reduzida, Royston em 1982 propôs uma modificação para que fosse possível o seu uso em amostras de dimensão superior e em 1999 facultou um algoritmo para amostras com dimensão $3 \leq n \leq 5000$, que tem vindo a apresentar bons resultados, (Razali & Wah, 2011).

- **Teste de *Kolmogorov–Smirnov* (KS):** O teste é usado para decidir se a distribuição da variável sob estudo, $F(X)$, numa determinada amostra, provém de uma população com distribuição específica, $F_0(X)$. Em particular, este teste é comummente usado para testar se a distribuição da variável segue ou não a distribuição normal, com parâmetros $\mu$ e $\sigma$, ou seja, objetiva-se testar:

  $$H_0 : X \sim N(\mu, \sigma) \qquad vs \qquad H_1 : X \nsim N(\mu, \sigma).$$

  Note-se que se estivéssemos a testar outro tipo de distribuição que não a normal as hipóteses seriam então $H_0 : F(X) = F_0(X) \quad vs \quad H_1 : F(X) \neq F_0(X)$.

  O cálculo da estatística de teste, inicia-se com a ordenação das observações da variável $X$ por ordem crescente, calculando em seguida a frequência acumulada de cada observação. Para obter a estatística de teste é necessário calcular a diferença entre a frequência acumulada de cada uma das observações e a frequência acumulada que essa observação teria se a sua distribuição de probabilidade fosse normal, assim como a mesma diferença relativamente à observação anterior. Assim, a estatística de teste consiste na maior destas duas diferenças, (Conover, 1999), isto é,

  $$D = \max\{\max(|F(x_i) - F_0(x_i)|) ; \max(|F(x_{i-1}) - F_0(x_i)|)\},$$

  em que $F_0(X) \sim N(\mu, \sigma^2)$.

  O valor crítico a partir do qual se efetua a comparação da estatística de teste explanada encontra-se tabelado na literatura ($Dtabela(\alpha)$), sendo que se rejeita a hipótese nula, sob o nível de significância $\alpha$, caso se tenha que $D_{obs} \geq Dtabela(\alpha)$. Em inúmeros *software* estatísticos o presente teste é implementado com base no calculo do $p - value$ isto é, o menor valor de $\alpha$ a partir do qual se tem que $D_{obs} \geq Dtabela(\alpha)$. Nesse sentido para um nível de significância $\alpha$, rejeita-se $H_0$ se $p - value \leq \alpha$.

  Note-se que a aplicação do teste de KS assume que os parâmetros populacionais $\mu$ e $\sigma$ são conhecidos. Não obstante, tal situação é pouco comum e assim, na maioria dos casos, a estatística $D$ não pode ser aplicada com rigor quando em vez dos verdadeiros valores de $\mu$ e $\sigma$ se conhecem apenas estimativas amostrais. Para corrigir este problema, foi proposta por Lilliefors em 1967 uma modificação deste teste, (Marôco, 2010).

- **Teste de KS com correção de *Lilliefors*:** Consiste num teste de normalidade baseado no teste de Kolmogorov – Smirnov como visto anteriormente, sendo útil em situações em que se pretende comparar a distribuição de frequências acumuladas das observações da variável com uma distribuição teórica, cujos parâmetros foram estimados a partir da amostra. Assim é comummente usado para testar a hipótese nula de que os dados resultam de uma população normalmente distribuída, quando a hipótese nula não especifica qual distribuição normal; isto é, não especifica os parâmetros populacionais da distribuição normal, $\mu$ e $\sigma$.

- **Teste de *Anderson-Darling* (AD):** De acordo com Arshad et al., este teste é o mais poderoso dos testes EDF. O presente teste objetiva testar a hipótese de que uma dada amostra tenha sido retirada de uma determinada população com função de distribuição acumulada contínua $F(x)$. Suponha-se que, $X_1, X_2, \cdots, X_n$ é uma amostra aleatória ordenada. Assim, as hipóteses traduzidas pelo referido teste são:

  $H_0$ : *a amostra tem distribuição* $F(x)$ *vs.* $H_1$ : *a amostra não tem distribuição* $F(x)$.

Anderson e Darling (1954) definiram a estatística de teste como:

$$W_n^2 = n \int_{-\infty}^{\infty} [F_n(x) - F^*(x)]^2 \psi(F^*(x)) dF^*(x),$$

onde $\psi$ é uma função de pesos não negativa que pode ser definida por

$$\psi = [F^*(x)(1 - F^*(x))]^{-1}.$$

Por forma a tornar a computação desta estatística mais fácil, pode-se aplicar a seguinte forma:

$$W_n^2 = -n - \frac{1}{n} \sum (2i-1) \{logF^*(X_i) + \log(1 - F^*(X_{n+1-i})\},$$

onde $F^*(X_i)$ é a função distribuição cumulativa de uma distribuição especifica; $X_i$'s são os dados ordenados e $n$ designa o tamanho da amostra, (Dufour et al., 1998; Leotti et al., 2005; Razali & Wah, 2011).

Para uma distribuição com parâmetros conhecidos, $\mu$ e $\sigma$, temos os valores da função de distribuição acumulada da estatística $W_n^2$ tabulados. Nessa tabela procura-se o valor de percentil estatístico crítico ou permitido, $W_{cr}^2$, para o tamanho da amostra n, para o nível de significância desejado ou requerido.

Se o valor calculado pela estatística de teste for inferior ao valor de percentil estatístico crítico da tabela, isto é, $W_{n,obs}^2 < W_{cr}^2$, então não se rejeita $H_0$, ou seja, não se rejeita que a amostra tenha distribuição $F(x)$, para um nível de significância $\alpha$. Caso contrário, se o valor calculado pela estatística de teste for igual ou superior ao valor de percentil estatístico crítico da tabela, isto é, $W_{n,obs}^2 \geq W_{cr}^2$, então a distribuição $F(x)$ deve ser rejeitada, ou seja, rejeita-se $H_0$, para o nível de significância $\alpha$.

**2. Autocorrelação dos Resíduos**

Pela hipótese $[H_3]$ já referida, sabe-se que os resíduos devem ser não autocorrelacionados. Isto é, não deverá existir correlação serial ao nível dos resíduos e a magnitude de um resíduo não deve influenciar a magnitude dos resíduos seguintes, (Marôco, 2010).

Note-se que assumindo normalidade $[H_4]$ a ausência de autocorrelação é equivalente à independência dos resíduos.

Para ser verificada empiricamente esta condição, recorre-se à construção de gráficos de resíduos vs valores preditos, que devem apresentar manchas de pontos aleatórios com o mesmo tipo de dispersão em torno do eixo das abcissas. Também se pode recorrer a testes de deteção de autocorrelação dos resíduos, entre os quais se destacam os testes de Durbin-Watson, Box-Pierce e Ljung-Box.

O **teste de Durbin-Watson** (DW) é um dos testes mais usados para a deteção de padrões de autocorrelação do tipo AR(1) e fundamenta-se numa estatística, designada pela letra $d$ ou

pelas iniciais DW, definida pelo quociente

$$d = DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=2}^{n} e_t^2} \approx 2(1 - \widehat{\rho}_{e_t;e_{t-1}}) = 2 - 2\widehat{\rho}_{e_t;e_{t-1}}, \tag{4.28}$$

em que $e_t$ se refere ao resíduo de estimação do modelo OLS e $\rho$ o coeficiente de correlação linear que tem de estar compreendido entre -1 e 1. Figura no numerador a soma de quadrados das diferenças entre resíduos de observações consecutivas, enquanto o denominador é, simplesmente, a soma de quadrados dos resíduos, e, por conseguinte, a estatística não poderá assumir valores negativos.

Tendo por base a equação 4.28 apresentada e atendendo a que $\hat{\rho}$ deverá variar entre -1 e 1, constata-se que $d$ toma valores entre 0 e 4. Se $d \approx 2$ é plausível inferir que não existe autocorrelação entre resíduos, na medida em que $\widehat{\rho}_{e_t;e_{t-1}} \approx 0$, isto é, não se deverá rejeitar a hipótese nula. A hipótese alternativa do teste é escolhida com base no valor da estatística de teste $d$. Assim, se $0 < d < 2$, tendo em conta que $0 < \widehat{\rho}_{e_t;e_{t-1}} < 1$, a hipótese alternativa testa autocorrelação positiva. Por outro lado, se $2 < d < 4$, tendo em conta que $-1 < \widehat{\rho}_{e_t;e_{t-1}} < 0$, a hipótese alternativa testa autocorrelação negativa entre os resíduos, (Oliveira et al., 2011).

Durbin e Watson enquadraram o valor crítico da distribuição dentro de dois limites, designados por $d_l$ e $d_u$, independentes da matriz X e dependentes apenas da sua ordem, i.e, de n e de k+1. Convém observar que se $d_l \leq DW \leq d_u$ ou se $4 - d_u \leq DW \leq 4 - d_l$ o teste é inconclusivo.

Note-se no entanto que se a decisão do teste for no sentido de assumir autocorrelação esta será gerada por um processo autoregressivo de ordem um. Este teste tem como desvantagem não ser capaz de testar outras estruturas de autoregressão ao nível dos erros.

Em suma, formulando matematicamente as hipóteses explanadas, o teste de Durbin-Watson pode ser expresso como

$$H_0 : \rho = 0 \text{ vs } H_1 : \rho > 0 \text{ ou } \rho < 1. \tag{4.29}$$

Outros testes alternativos ao teste referido são os **testes de Box-Pierce** e de **Ljung-Box**, tendo em conta que o segundo teste é mais generalizado que o primeiro, pois resulta da melhoria do anterior. E, por conseguinte, foi constatado na literatura que o teste de Ljung-Box permite obter resultados melhores do que o seu originário, (Ljung & Box, 1978). Porém, os dois testes assumem as seguintes hipóteses:

$$H_0 : \text{Os resíduos são i.i.d. vs } H_1 : \text{Os resíduos não são i.i.d.} \tag{4.30}$$

Quanto às estatísticas de teste para o teste de Box-Pierce e para o teste de Ljung-Box são, respetivamente

$$Q(K)_{\text{Box-Pierce}} = n\sum_{j=1}^{K}\widehat{\rho}_j^2 \text{ e } Q(K)_{\text{Ljung-Box}} = n(n-2)\sum_{j=1}^{K}\frac{\widehat{\rho}_j^2}{n-j}, \tag{4.31}$$

com $\hat{\rho} = \frac{\sum_{i=k+1}^{n} e_i e_{i-k}}{\sum_{i=1}^{n} e_i^2}$, onde $K$ é o número de desfasamentos que toma na função de autocorrelação estimada $\widehat{\rho}$ e $n$ o tamanho da amostra em estudo.

Deste modo, no caso da estatística de teste do teste de Box-Pierce, tem-se que

$$Q(K)_{\text{Box-Pierce}} \sim \chi^2_k;$$

enquanto que,

$$Q(K)_{\text{Ljung-Box}} \sim \chi^2_{k-p-q},$$

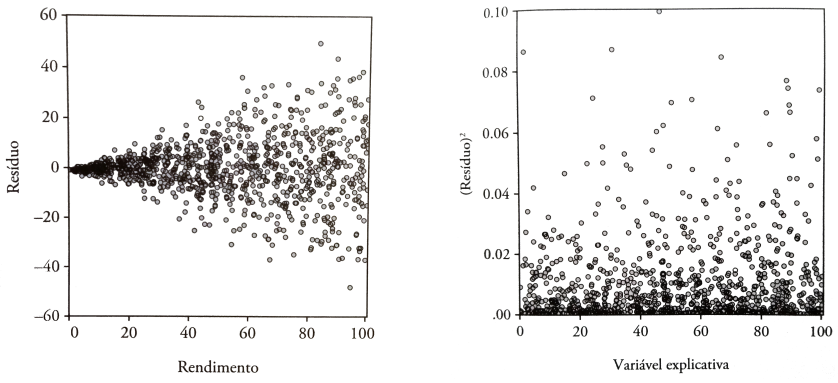
onde $p$ e $q$ representam as ordens do modelo ARMA (modelo autoregressivo de médias móveis), uma vez que este teste admite que a estrutura de autocorrelação seja explicada por um modelo ARMA(p,q), (McLeod & Li, 1983).

Assim sendo, a um nível de significância $\alpha$, no que refere ao teste de Box-Pierce rejeita-se a hipótese nula se $Q(K)_{\text{Box-Pierce}} > \chi^2_{1-\alpha,k}$, já no que respeita ao teste de Ljung-Box rejeita-se a hipótese nula se $Q(K)_{\text{Ljung-Box}} > \chi^2_{1-\alpha,k-p-q}$, (Lewis-Beck, 1993).

### 3. Heteroscedasticidade

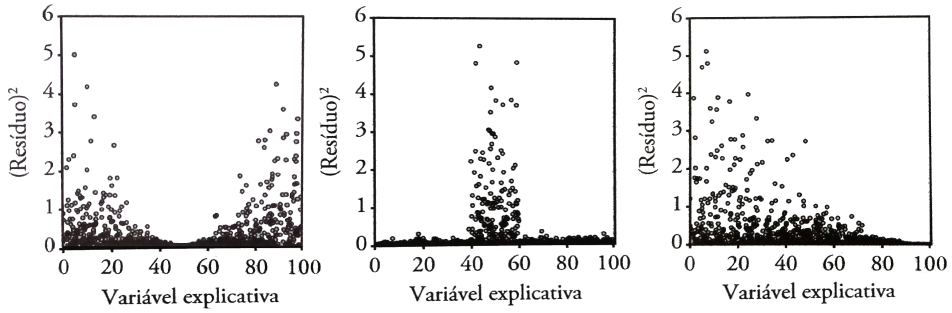
No modelo de regressão linear, admite-se que, a conjuntos diferentes de valores das variáveis explicativas corresponderão, em regra, médias da variável dependente diferentes. No entanto, a variância da variável dependente será a mesma. Decorre este último ponto de um pressuposto, o de uma variância das perturbações (e, por consequência, variância também da variável explicativa) constante, independente dos valores assumidos pelas variáveis explicativas, (Krzanowski, 1995). Esta hipótese diz respeito à hipótese já mencionada, $[H_2]$.

A hipótese de heteroscedasticidade (ou ausência de homocedasticidade) pode ser testada empiricamente numa primeira fase, recorrendo a gráficos entre os resíduos e os valores ajustados, tais como os apresentados nas figuras 4.4 e 4.5.




(a) Presença de heteroscedasticidade (b) Ausência de heteroscedasticidade

Figura 4.4: Heterocedasticidade




Fonte: (Oliveira et al., 2011)

Figura 4.5: Exemplos da presença de heteroscedasticidade

No gráfico 4.4a, a existência de um V na horizontal permite evidenciar a presença de heteroscedasticidade nos resíduos (é visível que à medida que a variável Rendimento toma valores mais altos, a variabilidade dos resíduos também aumenta), enquanto que no gráfico 4.4b estamos perante resíduos com ausência de heteroscedasticidade, pois os resíduos apresentam um comportamento completamente aleatório.

A figura 4.5, representa três possíveis situações onde é possível verificar a existência de heteroscedasticidade, na medida em que há um padrão para a distribuição dos resíduos.

Para além da análise empírica através de representação gráfica, o pressuposto da homogeneidade dos resíduos pode ser testado recorrendo a testes de hipóteses como o teste de Goldfeld–Quandt, o teste de Breusch–Pagan, o teste de White e o teste de Harrison–McCabe, (Lewis-Beck, 1993; Lim & Loh, 1996).

O **teste de Goldfeld-Quandt** é usado para testar a heteroscedasticidade dos resíduos, pressupondo que $\sigma_i^2$ se encontra positivamente correlacionado com uma das variáveis explicativas, por exemplo, através de $\sigma_i^2 = \sigma^2 X_{k,i}^2$ para a $k$-ésima variável explicativa. O presente teste começa por sugerir ordenar as observações por ordem crescente, de acordo com os valores de $X_k$ (variável independente suspeita de causar heteroscedasticidade).

Estabelece-se que $n$ representa o número de observações e $c$ designa o número de observações centrais omitidas para a realização do presente teste.

Este teste consiste em dividir as observações $n - c$ em dois subgrupos de igual tamanho. Estes subgrupos são constituídos de forma a que um contenha os valores mais reduzidos de $X_k$ e o outro os valores mais elevados de $X_k$. São ajustadas as regressões OLS separadamente, sendo posteriormente obtida a respetiva soma residual dos quadrados $SQR_1$ e $SQR_2$. Assumindo a normalidade dos erros e a homoscedasticidade, com SQR$\sim \chi_{df}^2$, em que $df = \frac{(n-c-2k)}{2}$, onde k é o número de parâmetros a serem estimados, incluindo o coeficiente independente.

A estatística de teste é dada por, (Rana et al., 2008):

$$GQ = \frac{SQR_2/df}{SQR_1/df} = \frac{SQR_2}{SQR_1} \sim F_{\frac{n-c-2k}{2},\frac{n-c-2k}{2}}. \tag{4.32}$$

Visto que este teste apenas considera que o padrão de heteroscedasticidade depende de uma única variável explicativa, uma generalização é o teste de Breusch-Pagan que permite que essa dependência inclua mais variáveis explicativas.

Assim, o **teste de Breusch–Pagan** constitui um avanço relativamente a metodologias mais antigas de deteção de heteroscedasticidade, que requeriam a identificação de uma única variável (comummente, uma das variáveis explicativas), (Marôco, 2010; Oliveira et al., 2011). O pressuposto subjacente ao teste de Breusch–Pagan é o de que os erros $e_t$ são independentes, com média nula e variância dada por:

$$\sigma_i^2 = h(\alpha_1 + \alpha_2 Z_{2i} + \alpha_3 Z_{3i} + ... + \alpha_p Z_{pi}), \tag{4.33}$$

onde $h(.)$ é função que não é necessário especificar, de uma combinação linear de variáveis observáveis $Z_2, Z_3, .., Z_p$, as quais designam os valores observados das variáveis que se julga relacionadas com a eventual heteroscedasticidade. Sobre as referidas variáveis é realizada uma nova regressão (regressão auxiliar).

Na existência da homoscedasticidade, tem-se que $\alpha_2 = \alpha_3 = ... = \alpha_p = 0$, em resultado tem-se que $\sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2 = h(\alpha_1)$, ou seja, uma constante. Assim, a hipótese nula do teste de Breusch-Pragan traduz-se pelo pressuposto da existência de homoscedasticidade nos resíduos:

$$H_0 : \alpha_2 = \alpha_3 = ... = \alpha_p = 0. \tag{4.34}$$

Se pelo contrário, algum dos coeficientes $\alpha_2, \alpha_3, ..., \alpha_p$ for diferente de zero, então $\sigma_i^2$ dependerá de $i$, denunciando uma violação da homoscedasticidade da variância dos resíduos. A estatística de teste é:

$$\frac{SQE}{2\tilde{\sigma}^4} \cong \chi^2_{(p-1)}, \tag{4.35}$$

onde SQE é a soma de quadrados explicada na regressão auxiliar e $\tilde{\sigma}^4 = (\tilde{\sigma}^2)^2 = (\frac{\sum e_i^2}{n})^2$. Um valor amostral da estatística de Breusch–Pagan maior do que o valor crítico, para $p-1$ graus de liberdade e para o nível de significância pretendido, levará à rejeição da hipótese nula e à conclusão pela presença de heteroscedasticidade, (Johnson, 81).

Uma generalização ainda maior dos testes referidos é o teste de White, isto porque não necessita de um padrão de heteroscedasticidade.

O **teste de White**, de entre as várias alternativas para testar a homogeneidade dos resíduos, é o de aplicação mais generalizada uma vez que não tem conjeturas quer sobre

a forma de heteroscedasticidade (heterogeneidade das variâncias dos resíduos) quer sobre a normalidade da distribuição dos erros, (Oliveira et al., 2011). Este teste estima pelo método OLS os parâmetros usando o modelo original e calcula os resíduos usando um modelo de regressão auxiliar considerando $e_i^2$ como variável dependente e como variáveis dependentes todas as variáveis independentes originais $(X_1, X_2, \ldots, X_k)$, os seus quadrados e produtos cruzados tomados duas a duas. Neste teste as hipóteses são:

$$H_0 : \text{As variâncias dos resíduos são homogéneas}$$
$$\text{vs}$$
$$H_1 : \text{As variâncias dos resíduos não são homogéneas.}$$

A estatística de teste é *(White*, 1980):

$$W = nR^2 \overset{a}{\sim} \chi^2_{a-1}, \tag{4.36}$$

onde $n$ é a dimensão da amostra, $a$ é o número de coeficientes da regressão auxiliar de White e $R^2$ é o coeficiente de determinação da regressão auxiliar.

Sob $H_0$, W tem distribuição $\chi^2_{a-1}$ e rejeita-se $H_0$ se $W_{obs} \geq \chi^2_{a-1}$, (Marôco, 2010). Assim, um valor observado da estatística W superior ao valor crítico, para o número de graus de liberdade indicado e o nível de significância pretendido, ditará a rejeição da hipótese nula e a conclusão pela presença de heteroscedasticidade.

Relativamente ao teste de White, o teste de Breusch–Pagan tem a vantagem de ser um teste construtivo, na medida em que incorpora uma hipótese específica sobre o padrão de heteroscedasticidade, suscetível de utilização na fase posterior de reestimação do modelo.

Note-se que os testes de White e de Breusch–Pagan têm apenas validade assintótica. Não há indicações seguras e com caráter de generalidade quanto à qualidade da aproximação em amostras de dimensão finita. Se os dois testes conduzem, no limite, a conclusão idêntica sobre a existência ou não de heteroscedasticidade, pode acontecer, em amostras de dimensão finita, que levem a conclusões contraditórias, (Oliveira et al., 2011).

O teste de Goldfeld-Quandt é um teste usualmente aplicado quando é possível ordenar as observações em termos da variância crescente do termo de erro (ou uma variável independente que supostamente causaria heteroscedasticidade), sendo o seu uso vantajoso na medida em que o teste não é sensível à suposição de normalidade dos resíduos, como o teste de Breusch–Pagan, (Dufour et al., 2004).

Um teste relacionado com os testes anteriormente explanados mas pouco utilizado, é o **teste de Harrison–McCabe** que foi proposto por Harrison e McCabe (1979) e se baseia unicamente nos resíduos de regressão obtidos por aplicação de OLS. A estatística de teste é dada por:

$$HM = \frac{e_1^T e_1}{e^T e}, \tag{4.37}$$

onde $e_1$ representa algum um subconjunto dos resíduos de regressão OLS, $e$, tal que sob a alternativa a variância é menor. A dimensão $T_1$ do subvetor $e_1$ é arbitrária, exceto se há conhecimento adicional relativamente à origem da heteroscedasticidade. E $T$ representa a dimensão do vetor $e$.

Sob $H_0$, a estatística HM deve ser próxima de $\frac{T_1}{T}$, sendo que se rejeita $H_0$ quando $HM$ é muito reduzido. Não obstante, o problema reside na dificuldade inerente à avaliação da distribuição de $HM$, em detrimento da estatística $GQ$ de Goldfeld-Quandt, o que justifica a utilização reduzida do teste de Harrison-McCabe na prática. Harrison e McCabe mostram que HM é delimitada pelas variáveis aleatórias $HM_L$ e $HM_U$ as quais dependem apenas de $T$, $T_1$, e sugerem um teste de limites semelhante ao teste de Durbin Watson.

O teste de Harrison-McCabe pode ser derivado do método dos multiplicadores de Lagrange, se a alternativa for devidamente definida. Por forma a comprovar o referido anteriormente, assume-se que a amostra é ordenada tal que $E(u_t^2) = \sigma_1^2$ para $t \leq T_1$, e $E(u_t^2) = \sigma_2^2$ para $t > T_1$. Além disso assume que $T_1$ é conhecido. Reparametrizando $\theta = \frac{\sigma_2^2 - \sigma_1^2}{\sigma_2^2}$, o teste anterior é equivalente a considerar

$$H_0 : \theta = 0 \quad \text{e} \quad H_1 : \theta > 0.$$

A matriz de covariância de perturbação é dada por

$$\sigma_2^2 V(\theta) = \sigma_2^2 diag(1-\theta, \ldots, 1-\theta, 1, \ldots, 1), \tag{4.38}$$

Tem-se $|V(\theta)| = (1-\theta)^{T_1}$, ou seja,

$$d(\theta) = -2^{-1}|V(\theta)|^{-1}\frac{\partial |V(\theta)|}{\partial \theta} = -2^{-1}(1-\theta)^{-T_1}T_1(1-\theta)^{T_1-1}. \tag{4.39}$$

Além disso,

$$V(\theta)^{-1} = diag((1-\theta)^{-1}, .., (1-\theta)^{-1}, 1, ..., 1), \tag{4.40}$$

ou seja,

$$A(\theta) = \frac{\partial V(\theta)^{-1}}{\partial \theta} = diag((1-\theta)^{-2}, .., (1-\theta)^{-2}, 0, ..., 0). \tag{4.41}$$

A derivada parcial de função de *log*-verosimilhança em função de $\theta$, avaliada nas estimativas de máxima verosimilhança restrita, $\beta = \hat{\beta}$, $\sigma^2 = \hat{\sigma}^2$, onde $\hat{\beta}$ e $\hat{\sigma}^2$ são novamente as estimativas

familiares OLS, é, portanto,

$$\begin{aligned}\frac{\partial L}{\partial \theta} &= d(0) - \frac{e^T A(0) e}{2\hat{\sigma}_2} \\ &= \frac{T_1}{2} - \frac{T}{2}\frac{e_1^T e_1}{e^T e} \\ &= \frac{T_1}{2}(1 - \frac{e_1^T e_1/T_1}{e^T e/T}).\end{aligned} \tag{4.42}$$

O termo entre parênteses na última expressão deve ser perto de zero quando $H_0$ é verdade, pelo que se rejeita $H_0$ sempre que 4.42 é positivo e muito grande, que equivale a rejeitar $H_0$ quando a estatística $HM$ de Harrison–McCabe assume valores reduzidos, (Krämer & Sonnberger, 1986).

Outra alternativa para testar a presença de heteroscedasticidade é dada pelo **teste de validação global das suposições de modelos lineares** (*Global validation of linear model assumptions*), cuja função no *software* R é *gvlma()*. O presente teste permite executar testes de hipóteses ao nível global do modelo linear, por forma a verificar quatro suposições inerentes ao mesmo. Assim, as quatro conjeturas indagadas pelo teste são:

- *Skewness* (Assimetria): a decisão, neste caso, baseia-se no teste da hipótese de que a distribuição dos erros é positiva ou negativamente enviesada, necessitando de transformação por forma a satisfazer os pressupostos da normalidade. A rejeição da hipótese nula, a partir de um valor de $p - value$ inferior a 0.05 (a um nível de significância de 5%), indica que se deve proceder a uma transformação dos dados;

- *Kurtosis* (Curtose): Neste ponto é testado se a distribuição dos erros possui curtose demasiado elevada (com picos demasiadamente pronunciados) e se é necessário alguma transformação de modo a satisfazer os pressupostos da normalidade. A rejeição da hipótese nula com $p - value$ inferior 0.05 indica que se deve proceder a transformação dos dados;

- *Link Function* (Função de ligação): A questão colocada consiste no teste da adequação da aplicação de regressão linear categórica ou se em alternativa será mais adequado a aplicação de um modelo linear generalizado, em particular regressão logística ou binomial. A obtenção de valores $p-value$ inferiores a 0.05 rejeitam a adequação de regressão linear aos dados em estudo;

- *Heteroscedasticity* (heteroscedasticidade): o presente teste verifica se a variância dos resíduos do modelo é constante. A estatística de teste inerente é

$$\left(\frac{1}{\sqrt{2\widehat{\sigma}_V^2 n}} \sum_{i=1}^{n} (V_i - \overline{V})(R_i^2 - 1)\right)^2,$$

  onde $V$ representa o vetor $n \times 1$, tal que $V = (1, 2, \ldots, n)^t/n$, $n$ faz referência ao número de observações do modelo e $R$ define os resíduos estandardizados. Com $\widehat{\sigma}_V^2 =$

$\frac{1}{n}\sum_{i=1}^{n}(V_i - \overline{V})^2$. A rejeição da hipótese nula, quando o $p - value$ é inferior a 0.05 indica que os resíduos possuem heteroscedasticidade e, por conseguinte, que a variância dos erros não é constante, (Peña & Slate, 2006).

Para além das hipóteses apresentadas anteriormente é ainda possível avaliar de forma global o modelo de regressão linear na medida em que o teste apresenta uma componente relativa a avaliação global, sendo que basta que uma das hipóteses supracitadas falhe para que o modelo seja rejeitado.

**4. Ortogonalidade entre as variáveis independentes**

A colinearidade consiste na forte correlação que as variáveis independentes exercem entre si. Quando esta condição está presente a análise do modelo de regressão pode não ser coerente ou até desprovida de significado (inclusivé os coeficientes estimados pelo método OLS podem não ser obtidos devido à matriz $(X^TX)$ não ser invertível).

Quando se está perante situações ótimas, as variáveis independentes são ortogonais, quer dizer, não estão correlacionadas, o modelo ajustado e os coeficientes de regressão podem usar-se com objetivos inferenciais e de estimação, (Alin, 2010; Brown, 1993; Oliveira et al., 2011).

A condição de colinearidade pode ser averiguada por várias técnicas, sendo seguidamente apresentadas algumas delas:

- É efetuada a análise da matriz de correlações, em que os coeficientes de correlação resultantes devem ser reduzidos para que assim a colinearidade não exista, contudo esta técnica só consegue analisar a colinearidade das variáveis duas a duas, (Alin, 2010; Marôco, 2010).

- Outra técnica que consegue verificar a colinearidade é a utilização do VIF – *Variance Inflation Factor* (fator de inflação da variância) – dado por $(\frac{1}{1-R_i^2})$. Dado que a variância dos coeficientes de regressão é conseguida através da seguinte expressão:

$$Var(\widehat{\beta}_i) = \sigma^2(\frac{1}{1-R_i^2}) \times \frac{1}{\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2}, \tag{4.43}$$

  sendo $R_i^2$ o coeficiente de determinação da regressão auxiliar da variável explicativa $X_i$ relativamente às restantes.

  Quando o coeficiente de regressão desta regressão auxiliar é nulo, a $Var(\widehat{\beta}_i)$ é a menor possível, pelo que a correlação é nula. Caso contrário, quanto maior for a colinearidade entre a variável $X_i$ e as restantes variáveis independentes, maior é o valor da variância. A análise aos valores de VIF para um valor acima de 10 indica que a colinearidade nas variáveis independentes pode estar a influenciar as estimativas de $\beta_i$, (Marôco, 2010; Miles, 2005).

- Sendo esta técnica de aferir a colinearidade relacionada com a anterior, calcula-se a tolerância da variável $X_i$, determinada a partir de $T = 1 - R_i^2 = \frac{1}{VIF}$. Posto que, valores de $T$ baixos, próximos de zero, significam que a variável $X_i$ pode ser escrita através de

outras variáveis independentes e, por conseguinte, a colinearidade confirmada. (Miles, 2005)

Posteriormente à validação destes pressupostos, a interpretação referente ao modelo de regressão pode ser realizada de modo mais confiável.

## 4.3 Transformação de *Box-Cox*

As *Power Transformations* designam um conjunto de técnicas comummente aplicadas na modificação de variáveis, usadas nos casos em que os pressupostos assumidos pelo modelo de regressão linear referidos na secção 4.2.8 não são válidos no conjunto de dados ou nos resíduos, tendo como objetivo a validação desses pressupostos e a transformação da não linearidade dos modelos numa expressão de forma linear, para se usar a teoria estatística associada a modelos de regressão linear, (Weisberg, 2001). Inserida nestas transformações, destaca-se a transformação de Box-Cox, nomeadamente quando o pressuposto de normalidade não é validado, cuja autoria pertence aos estatísticos George Box e David Cox em 1964, (Box & Cox, 1964; Li, 2005). No entanto, esta transformação também é eficiente no caso em que os outros pressupostos não sejam válidos, concretamente na falha da homogeneidade dos resíduos, isto é, quando os resíduos contêm heteroscedasticidade.

Desta forma, a transformação de Box-Cox é dada por, (Yeo & Johnson, 2000):

$$x_i^{\lambda} = \begin{cases} \frac{x^{\lambda}-1}{\lambda}, & \text{para } \lambda \neq 0 \\ log(x), & \text{para } \lambda = 0 \end{cases} \tag{4.44}$$

onde $x$ representa os valores da variável sobre os quais se pretende aplicar a transformação e o valor de $\lambda$ a potência à qual os dados de uma variável devem ser elevados.

O procedimento consiste na identificação do expoente $\lambda$ que resulte na melhor transformação dos dados numa distribuição Normal. Concretamente, o algoritmo procura o valor de $\lambda$ entre -5 e 5 até o melhor valor ser encontrado.

As transformações recorrentes são: $x^{-2}$ ($\lambda = -2$), $x^{-1}$ ($\lambda = -1$), $x^{-0.5} = 1/\sqrt{x}$ ($\lambda = -0.5$), $x^0 = log(x)$ ($\lambda = 0$), $x^{0.5} = \sqrt{x}$ ($\lambda = 0.5$), $x^1$ ($\lambda = 1$) e $x^2$ ($\lambda = 2$).

É de salientar que a presente transformação apenas é possível usar quando os dados são positivos e diferentes de zero. Contudo, quando tais situações não se verificam é possível adicionar um valor constante C a todos os dados, permitindo assim que todos se tornem positivos antes da aplicação da transformação, ou seja, modificando-os de $x$ em $x^* = (x+C)^{\lambda}$, (Sakia, 1992).

Para verificar a eficiência da transformação implementada, recorre-se à análise dos dados transformados tal como explanado na secção 4.2.8.

## 4.4 Métodos de estimação na presença de autocorrelação

Perante a existência de autocorrelação, existem métodos que permitem estimar os coeficientes de regressão tendo em conta a autocorrelação existente. Nesta secção serão abordados alguns desses métodos. Os métodos em análise supõem que as perturbações são geradas por um processo de autocorrelação de 1ª ordem, AR(1), com o coeficiente de autocorrelação de 1ª ordem, $\rho$, desconhecido.

O primeiro método de estimação apresentado é o **método de Cochrane-Orcutt**, que se baseia no método OLS, (Verbeek, 2004). Este método propõe que se encontrem estimativas para os parâmetros $\beta_0, \beta_1, ..., \beta_k, \rho$ por forma a que se minimize a soma dos quadrados dos resíduos do modelo transformado, ou seja, obter $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k, \hat{\rho}$ de modo a que

$$\min_{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k, \hat{\rho}} S(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k, \hat{\rho}), \tag{4.45}$$

em que S(.) representa a soma dos quadrados dos resíduos tal que

$$S(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k, \hat{\rho}) = \sum_{i=2}^{n} \Big[ \big(Y_i - \hat{\rho} Y_{i-1}\big) - \big(\hat{\beta}_0(1-\hat{\rho}) + \hat{\beta}_1(X_{1,i} - \hat{\rho} X_{1,i-1}) + ... + \hat{\beta}_k(X_{k,i} - \hat{\rho} X_{k,i-1})\big)\Big]^2.$$

Denotando $Y_i^* = Y_i - \hat{\rho} Y_{i-1}$ e $\hat{Y}_i^* = \hat{\beta}_0(1-\hat{\rho}) + \hat{\beta}_1(X_{1,i} - \hat{\rho} X_{1,i-1}) + ... + \hat{\beta}_k(X_{k,i} - \hat{\rho} X_{k,i-1}$, a equação 4.45 pode ser reescrita como

$$\min_{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k, \hat{\rho}} \sum_{t=2}^{n} \big(Y_i^* - \hat{Y}_i^*\big)^2.$$

O referido método é também designado pelo método de Cochrane-Orcutt bietápico, visto que se esquematiza em dois passos, que são descritos seguidamente, (Oliveira et al., 2011):

- 1º Passo: Pelo método OLS estimar o modelo original

  $$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \varepsilon_i \quad i = 1, 2, ..., n,$$

  obter a série dos respetivos resíduos de estimação

  $$e_i^{(1)} = Y_i - \hat{\beta}_0^{(1)} - \hat{\beta}_1^{(1)} X_{1i} - ... - \hat{\beta}_k^{(1)} X_{ki}, \quad i = 1, 2, ..., n$$

  e por fim determinar $\hat{\rho}^{(1)}$, isto é, determinar uma estimativa de $\rho$, a partir da estimação por OLS do seguinte modelo

  $$e_i^{(1)} = \rho e_{i-1}^{(1)} + v_i, \quad i = 2, 3, ..., n,$$

  onde $v_i$ designa uma perturbação aleatória. Assim, o estimador OLS de $\rho$ é determinado da seguinte forma

  $$\hat{\rho}^{(1)} = \frac{\sum_{i=2}^{n} e_i^{(1)} e_{i-1}^{(1)}}{\sum_{i=2}^{n} {e_{i-1}^{(1)}}^2}.$$

- 2º Passo: Construir as variáveis transformadas com recurso à estimativa de $\rho$ obtida no 1º passo, conseguindo assim, o seguinte modelo transformado

$$\underbrace{\left(Y_i - \hat{\rho}^{(1)} Y_{i-1}\right)}_{Y_i^*} = \underbrace{\beta_0(1-\hat{\rho}^{(1)})}_{\beta_0^*} + \beta_1 \underbrace{\left(X_{1i} - \hat{\rho}^{(1)} X_{1,i-1}\right)}_{X_{1i}^*} + ... + \beta_k \underbrace{\left(X_{ki} - \hat{\rho}^{(1)} X_{k,i-1}\right)}_{X_{ki}^*} + \varepsilon_i - \hat{\rho}^{(1)} \varepsilon_{i-1}, \quad i = 2, 3, ..., n.$$

O modelo alcançado é estimado novamente com recurso ao OLS, obtendo-se diretamente $\hat{\beta}_0^{*(2)}, \hat{\beta}_1^{(2)}, ..., \hat{\beta}_k^{(2)}$, onde $\hat{\beta}_0^2$ é dado pela seguinte equação

$$\hat{\beta}_0^2 = \frac{\hat{\beta}_0^{*(2)}}{1 - \hat{\rho}^{(1)}}.$$

Este processo é repetido tantas vezes quantas as necessárias para se aproximar do mínimo da função $S(.)$ desejado, atingido por algum critério de convergência, Oliveira et al. (2011). Este é um método de estimação iterativo.

O segundo método abordado é o método de Durbin, sendo que este método recorre às diferenças generalizadas de 1ª ordem. Essas diferenças resultam do seguinte processo, em que se considera o modelo

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \varepsilon_i$$

e o mesmo se reescreve para o período anterior,

$$Y_{i-1} = \beta_0 + \beta_1 X_{1,i-1} + ... + \beta_k X_{k,i-1} + \varepsilon_{i-1}.$$

Seguidamente multiplicam-se ambos os membros da anterior equação por $\rho$ que será sempre diferente de zero (pois se for igual a zero não existe autocorrelação) e subtrai-se à equação do modelo original, resultando a seguinte expressão

$$\left(Y_i - \rho Y_{i-1}\right) = \beta_0(1-\rho) + \beta_1(X_{1i} - \rho X_{1,i-1}) + ... + \beta_k(X_{ki} - \rho X_{k,i-1}) + \varepsilon_i - \rho\varepsilon_{i-1}, \quad i = 2, 3, ..., n.$$

A equação anterior pode-se traduzir na seguinte:

$$Y_i = \beta_0^* + \rho Y_{i-1} + \beta_1 X_{1i} + \beta_1^* X_{1,i-1} ... + \beta_k X_{ki} + \beta_k^* X_{k,i-1} + u_i, \tag{4.46}$$

em que $u_i = \varepsilon_i - \rho\varepsilon_{i-1}$ representa o termo de perturbações aleatório, $\beta_j^* = -\rho\beta_j$ e $\beta_0^* = \beta_j(1-\rho)$, com $j = 2, 3, ..., k$.

O **método de Durbin** de estimação dos coeficientes de regressão de um modelo AR(1), segue os seguintes passos:

- 1º passo: Pelo método OLS estimar modelo dado na equação 4.46. Note-se que uma estimativa do coeficiente do termo $Y_{i-1}$ produz uma estimativa de $\rho$.
- 2º passo: Construir as variáveis transformadas a partir da estimativa para $\rho$, conseguida no passo anterior, para assim aplicar a estimação por OLS ao modelo transformado,

$$\underbrace{\left(Y_i - \hat{\rho}Y_{i-1}\right)}_{Y_i^*} = \underbrace{\beta_0(1-\hat{\rho})}_{\beta_0^*} + \beta_1 \underbrace{\left(X_{1i} - \hat{\rho}X_{1,i-1}\right)}_{X_{1i}^*} + ... + \beta_k \underbrace{\left(X_{ki} - \hat{\rho}X_{k,i-1}\right)}_{X_{ki}^*} + v_i, \quad i = 2, 3, ..., n.$$

  com $v_i$ uma perturbação.

O método descrito tem como vantagem a simplicidade inerente à sua aplicação face ao anterior, por este incluir apenas duas estimações de mínimos quadrados enquanto que o anterior exige o cálculo de um extremo de uma função. Porém, o método de Durbin tem como desvantagem a possibilidade de se obter um $\rho$ fora do intervalo de permitido, sendo esse intervalo de -1 a 1, como já foi referido na secção 4.2.8. Acrescida a esta vem a desvantagem de que existe a possibilidade de não se conseguir obter a regressão linear auxiliar de uma amostra de dimensão $n$, por causa do número elevado de regressores que ela pode conter, pois esta estimação só é possível se $n - 1 > 2k$, com $k$ o número de coeficientes. Por fim, tem como desvantagem a perda de eficiência na estimação da regressão auxiliar no processo de estimação dos coeficientes $\beta_1^*, \beta_2^*, ..., \beta_k^*$ não tendo em conta que $\beta_j^*$ é igual a $-\rho\beta_j$, $j = 2, 3, ..., k$.

## 4.5 Matriz de variâncias e covariâncias consistentes na presença de heteroscedasticidade e/ou autocorrelação

Nos últimos 20 anos, foram desenvolvidos inúmeros procedimentos por forma a estabelecer estimadores consistentes quer na presença de heteroscedasticidade (HC – *heteroskedasticity consistent*) quer na presença de heteroscedasticidade e autocorrelação (HAC – *heteroskedasticity and autocorrelation consistent)*, situações bastante comuns em análises econométricas.

Na sequência da equação 4.12 a matriz $\psi$ de variâncias e covariâncias é usualmente denotada por

$$\psi = VAR[\widehat{\beta}] = (X^T X)^{-1} X^T \Omega X (X^T X)^{-1}, \tag{4.47}$$

onde $\Omega = VAR[\varepsilon]$.

Para a realização de inferência estatística com base no modelo de regressão linear é essencial a existência de um estimador consistente para a matriz de covariâncias $\psi$. O tipo de estimador escolhido para $\psi$ depende dos pressupostos sob a matriz de covariância dos erros $\Omega$. Assim, tal como explanado anteriormente, no modelo de regressão linear clássico os erros assumem-se como sendo independentes e homocedásticos com variância $\sigma^2$ donde surge $\Omega = \sigma^2 I_n$ e

$\psi = \sigma^2(X^TX)^{-1}$, os quais podem ser estimados de forma consistente a partir do estimador OLS usual

$$\widehat{\sigma}^2 = (n-k-1)^{-1}\sum_{i=1}^{n} e_i^2, \tag{4.48}$$

onde $n$ é o número de observações e $k$ o número variáveis preditoras. No entanto, se o pressuposto de independência e/ou homocedasticidade for violado, o estimador $\widehat{\psi} = \widehat{\sigma}^2(X^TX)^{-1}$ é enviesado e a inferência baseada nesse estimador é comprometida (ou não será válida). Os estimadores HC e HAC permitem colmatar este problema a partir da estimativa de $\widehat{\Omega}$, a qual é consistente na presença quer de heteroscedasticidade quer de autocorrelação, (Zeileis, 2004).

Caso os erros sejam independentes, mas heterocedásticos, a sua matriz de covariância $\Omega$ é diagonal apesar dos elementos sobre a diagonal não serem constantes. Assim, foram desenvolvidos inúmeros estimadores HC os quais calculam a nova matriz de covariâncias dos estimadores de regressão $\widehat{\psi}_{HC}$.

Por outro lado, se os erros $\varepsilon_i$ para além de heteroscedásticos, não forem independentes, $\Omega$ não é diagonal e dificilmente é possível estimar a matriz de covariância dos erros. O método de estimação HAC pressupõe que os erros são heterocedásticos e autocorrelacionados estimando $\Phi = k^{-1}X^T\Omega X$, que representa essencialmente a matriz de covariâncias das funções de estimação. O procedimento dos estimadores HAC baseia-se em:

$$\Phi = \frac{1}{k}\sum_{i,j=1}^{k} w_{|i-j|}\widehat{V}_i\widehat{V}_j^T, \tag{4.49}$$

onde $w = (w_0, ..., wk-1)^T$ é um vetor de pesos. Os valores estimados pelas funções de regressão são dados por $V_i(\beta) = x_i(y_i - x_i^T\beta)$.

A titulo conclusivo, pode-se dizer que as metodologias robustas explanadas alteram apenas os valores dos erros padrão, das estimativas de teste e dos valores do $p-value$ associados aos coeficientes de regressão linear. De facto, as alterações aplicadas por este método verificam-se apenas ao nível dos valores supracitados, sendo que os restantes valores associados ao modelo de regressão obtidos, nomeadamente os valores dos coeficientes de regressão e do erro quadrático médio por aplicação do método dos mínimos quadrados (OLS) se mantêm inalterados.

As metodologias HC e HAC encontram-se disponíveis no pacote *sandwich* implementado no *software* R, o qual disponibiliza as funções *vcovHC* e *vcovHAC* responsáveis pela aplicação dos estimadores HC e HAC, respetivamente.

Os estimadores HAC encontram-se disponíveis não só para metodologias de regressão linear, mas também para modelos de regressão linear generalizados (*generalized linear models – glm*) e modelos de regressão robusta.

## 4.6 Deteção de *outliers* e observações influentes

Um *outlier* pode ser definido como sendo uma observação que parece ser inconsistente num conjunto de dados, isto é, uma observação atípica em comparação com as restantes observações desse conjunto de dados, (Barnett & Lewis, 1974).

O estudo de *outliers* é deveras importante, nomeadamente aquando da aplicação de regressão a um conjunto de dados, isto porque os *outliers* podem afetar os valores dos coeficientes de regressão estimados, e, por conseguinte, desvirtuar o modelo de regressão ajustado.

Para além de determinar os *outliers*, também é possível analisar se as observações são consideradas observações influentes, isto é, se influenciam o ajustamento do modelo.

Uma das estratégias de diagnóstico para a deteção de *outliers* e, por conseguinte, observações influentes consiste no estudo do comportamento dos resíduos. Considerando este estudo podem-se apresentar as seguintes alternativas:

- Recorrer ao *box-plot* dos resíduos;

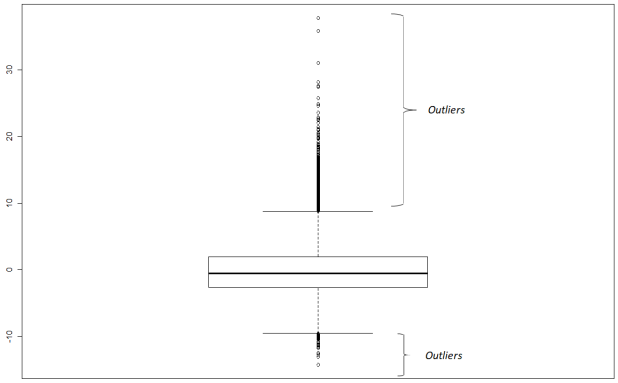



Figura 4.6: Exemplo do *box-plot* dos resíduos

- Calcular os resíduos studentizados (externos) ou, também, designados resíduos de Pearson estandardizados $r_j$, em função dos resíduos estandardizados $e_j$ e da *leverage* $h_j$, dados por

$$r_j = \frac{e_j^{'}}{\sqrt{1-h_j}} \sim t_{n-k-2}, \quad j = 1, ..., n, \tag{4.50}$$

onde $e_j^{'}$ têm variância constante e igual a 1 e $h_j$ é o elemento da diagonal da matriz do chapéu definida por $H = (X^T X)^{-1}(X^T)$.

Considerando amostras de elevada dimensão este valor deve estar compreendido entre -1.96 e 1.96 para um nível de significância de $\alpha = 0.05$, isto é, para um quantil de ordem 0.975 de $N(0,1)$ tem-se $Z_{0.975} = 1.96$. Se não estiverem compreendidos entre esses valores as observações podem considerar-se *outliers*. Em detalhe, esta alternativa passa pela

construção de gráficos de comparação de quantil para resíduos studentizados, considerando a distribuição t ou normal, nomeadamente QQ-plot para resíduos studentizados;

Após a aplicação das estratégias anteriores para a deteção de *outliers*, segue a aplicação de novas estratégias para verificar se temos observações influentes. Um ponto influente pode ser um outlier ou ter uma *leverage* elevada, isto é, ter grande efeito na estimação dos coeficientes de regressão.

Um ponto influente pode ser um outlier ou ter leverage elevada ou ambos. A distância de Cook é uma medida que combina estas duas propriedades com o objetivo de avaliar a influência de uma observação sobre o modelo. É dada por

$$DC_j = r_j^2 \frac{h_j}{1-h_j}. \tag{4.51}$$

Cook e Weisberg (1982) sugeriram que se esse valor fosse maior que um, era motivo de preocupação.

A regra para considerar que uma observação é influente é dada pela seguinte equação:

$$DC_j > \frac{4}{n-k-1},$$

onde $n$ é o número de observações e $k$ o número de regressores, (Chatterjee & Hadi, 2012). Porém, existem diferentes opiniões relativamente à escolha do ponto de corte (também chamado *threshold*), nomeadamente $DC_j > 1$ ou $DC_j > 4/n$.

Graficamente é possível representar as distâncias de Cook e o respetivo ponto de corte, por forma a facilitar a deteção das observações influentes, a figura 4.7 é um exemplo dessa representação gráfica, em que os pontos acima da reta (ponto de corte) são considerados observações influentes.

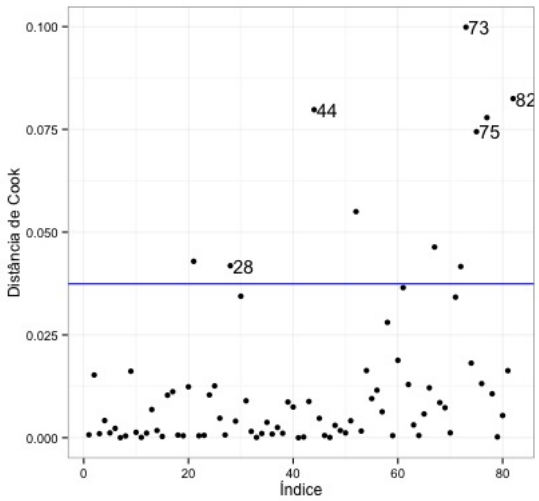



Figura 4.7: Exemplo da presença de observações influentes

## 4.7 Abordagem robusta

Em situações em que os pressupostos abordados na secção 4.2.8 não são validados ou quando se pretende reduzir a influência das observações atípicas mantendo-as no processo de estimação, é necessário recorrer a outras metodologias na implementação da modelação. Os métodos da regressão robusta têm como objetivo principal, atenuar o efeito do afastamento das hipóteses do modelo.

De facto, os métodos de estimação habituais (tal como o método dos mínimos quadrados ordinários, OLS), têm boas propriedades quando se aplicam nas condições previstas e quando os pressupostos são válidos. De acordo com o Teorema de Gauss-Markov: verificadas as hipóteses clássicas do modelo de regressão linear, os estimadores OLS de $\beta$ são os estimadores de variância mínima, na classe dos estimadores cêntricos de $\beta$ e lineares em $Y$. Os estimadores OLS são chamados estimadores BLUE (*Best Linear Unbiased Estimators*) - os melhores estimadores lineares e centrados (melhores no sentido de serem os mais eficientes ou terem menor variância, dentro da classe dos estimadores cêntricos e lineares). No entanto, podem ser desastrosos caso se verifiquem desvios nos pressupostos. Nesse sentido é importante utilizar técnicas que sejam robustas, de modo a que a existência de pequenos afastamentos relativamente às hipóteses do modelo não levem à produção de estimativas relativamente afastadas do que era previsto.

A aplicação de técnicas de análise estatística robusta é considerada um complemento às técnicas clássicas de mínimos quadrados. Isto deve-se ao facto de que se conseguem alcançar resultados semelhantes quando existe uma relação linear entre as variáveis com os erros normalmente distribuídos, porém diferem das técnicas clássicas no que refere a erros que não são normalmente distribuídos ou quando se está na presença de *outliers*, (Ryan, 2008).

Deve-se ter em conta que:

- Em situações em que os erros são normalmente distribuídos, e o conjunto de dados não possui erros nem valores *outliers*, a performance da regressão robusta deve ser próxima da do método dos mínimos quadrados (sendo que nesses casos o mais adequado é aplicar OLS);
- Quando as condições inerentes aos erros de regressão não são satisfeitas, deve permitir a obtenção de melhores resultados do que a aplicação do método do OLS;
- Não deve ser difícil de implementar computacionalmente nem de entender;
- A regressão robusta deve disponibilizar metodologias para construir intervalos de confiança e testes de hipóteses.

De uma forma genérica a regressão robusta atua de modo a atribuir pesos às observações, consoante a influência que exercem no processo de estimação. Desta forma, às observações cujos resíduos possuem elevada magnitude são atribuídos menores pesos. Ao limitar a influência das observações atípicas, a regressão robusta realiza um ajustamento que reflete principalmente a contribuição da maioria dos dados.

Existem inúmeros métodos de estimar os coeficientes associados aos métodos robustos. Na presente seção, faz-se referência aos estimadores-M e aos estimadores-MM que são estimadores robustos, (J. Branco & Pires, 2007).

Os estimadores-M usados na regressão robusta foram introduzidos por Huber (1973) e podem ser entendidos como uma extensão dos estimadores de mínimos quadrados.

A classe de M-estimadores foi estendida para todas as distribuições de probabilidade e generaliza o método da máxima verosimilhança, produzindo estimadores consistentes e assintoticamente normais (Heritier et al., 2009). Esta classe tem como objetivo minimizar a soma de uma certa função dos erros aleatórios, definida pela seguinte equação

$$\min_{\beta} \sum_{i=1}^{k} \rho(e_i), \tag{4.52}$$

onde $\rho(e_i)$ é uma função simétrica dos resíduos.

Mais tarde, foram desenvolvidos os estimadores-GM que permitiram melhorar os estimadores – M pelo facto de estes não serem resistentes a *outliers*.

Recentemente foram propostos os estimadores-MM por Yohai (1987), que são considerados, atualmente, uma da melhores opções dentro dos estimadores robustos. Os estimadores-MM da regressão são estimadores-M calculados a partir de estimativas iniciais convenientes. Assim, um estimador-MM corresponde ao mínimo local da função objetivo de um estimador-M, (Ryan, 2008).

Por forma a calcular os estimadores-MM robustos, computacionalmente, faz-se uso da função *rlm()* disponível no package MASS do *software R*.

Outros estimadores mais recentes e considerados os melhores atualmente, são os estimadores RR–MM que combina regressão ridge com estimadores MM, porém neste estudo não são abordados, (Maronna, 2011).

# Capítulo 5

# Enquadramento prático

O procedimento que antecede a implementação de análises Estatísticas dos dados consiste na exploração dos dados, a partir de ferramentas gráficas e análises exploratórias adequadas. Este procedimento é primordial na identificação e caracterização do conjunto de dados, em particular no que diz respeito a padrões existentes ou relações entre as variáveis. De facto, é possível adequar as análises de inferência estatística implementadas aos padrões existentes, bem como às características inerentes aos mesmos, permitindo que as metodologias aplicadas, e, por conseguinte, as conclusões inferidas sejam o mais úteis possível.

No presente capítulo são aplicadas, primeiramente, metodologias de análise exploratória com o objetivo de adequar metodologias estatísticas e selecionar variáveis do conjunto de dados. Posteriormente, objetivando-se dar resposta ao problema em estudo, procedeu-se à aplicação de regressão linear múltipla para que dessa forma fossem verificados quais os fatores que mais influenciam o Acréscimo Médio Anual projetado aos 12 anos (AMA_U_Proj12).

Note-se que, apesar de algumas variáveis estarem associadas ao ano de medida, nesta análise não é possível usar dados em painel, o que se deve ao facto de que as parcelas avaliadas em cada ano podem não coincidir.

Os dados foram processados usando o *software* QGIS. A análise estatística descritiva e inferencial foi realizada com auxilio do *software* R (versão 3.4.2) e do *software* E.Views (versão 10 SV).

## 5.1 Descrição do problema

O problema proposto pela entidade de acolhimento, RAIZ, consiste na definição de uma metodologia para integrar risco e incerteza na estimativa do crescimento e produção florestal, aquando da presença da praga que mais prejuízo causa atualmente na indústria papeleira: o

*Gonipterus platensis*, mais comummente designado por gorgulho-do-eucalipto.

Tendo por base o aumento da ocorrência de pragas, bem como doenças ou o elevado risco de incêndio, cujo aumento substancial se tem verificado atualmente, é primordial tentar avaliar o risco, modelando-o, para que sejam implementadas medidas de prevenção, minimizando as consequências e danos inerentes às referidas situações. De facto, as situações supracitadas traduzem-se em maior incerteza na obtenção de estimativas de produtividade e na tomada de decisão florestal, pelo que o estudo da forma como o risco e a incerteza podem ser integrados na modelação da produtividade, é profícuo, permitindo minimizar as consequências nefastas resultantes das ocorrências mencionadas. Porém o presente estudo recai apenas sobre a modelação do risco e incerteza na presença do *Gonipterus platensis*, isto é, de que forma a ocorrência da praga, o gorgulho do eucalipto, poderá afetar a estimativa da produtividade florestal em regiões de risco moderado a muito forte.

## 5.2 Descrição do conjunto de dados

No sentido de dar resposta ao problema colocado, fez-se uso de dois conjuntos de dados: dados relativos ao inventário florestal, e dados relativos à temperatura.

O primeiro conjunto de dados foi fornecido pela entidade de acolhimento, RAIZ. Este conjunto era formado inicialmente por 24904 observações e 42 variáveis. Não obstante, uma análise ao mesmo permitiu detetar a existência de observações com valores em branco ao nível de variáveis de relevo nos estudos subsequentes. Com efeito, foram removidas as observações cujas variáveis possuíam valores em branco nas variáveis AMA (foram eliminadas apenas 2 observações com esta característica) e nível de ataque existente nas parcelas (sendo a remoção inerente a este critério de maior dimensão, sendo que o conjunto de dados resultante era formado por 12648 observações), visto serem ambas variáveis preponderantes na presente investigação. No que se refere à variável relativa ao nível de ataque do *Gonipterus platensis*, verificou-se que algumas observações possuíam como codificação "não avaliadas". Assim, sendo o cerne da presente investigação avaliar a presença da espécie invasora em Portugal Continental, modelando o risco da sua presença na produtividade das parcelas, não se considerou útil usar as observações cuja variável referente à praga possuísse a codificação supracitada. Tal como é possível concluir a partir do anexo B.1 as parcelas com esta codificação apresentam motivos diversificados que justificam a sua não avaliação. A última remoção das parcelas com a codificação referida permitiu obter um conjunto de dados com as 12096 observações utilizadas no presente estudo.

Relativamente às variáveis utilizadas, o conjunto de dados referente ao inventário florestal era originalmente formado por 83 variáveis, das quais algumas possuem informação que se destina apenas à confirmação de valores medidos, pelo que, não acrescentando informação adicional ao conjunto, foram removidas. A lista das 83 variáveis originais é detalhadamente descrita no anexo A.

O segundo conjunto de dados possui informação relativa às temperaturas existentes nos locais das parcelas em estudo, as quais foram recolhidos a partir do site *worldweatheronline*.

Estes dados são referentes aos municípios que estão presentes no conjunto anterior, durante o período de 2011-2015, comum a ambos conjuntos de dados. Assim, este conjunto de dados contém 277 observações, referentes aos municípios onde se localizam as parcelas em estudo do primeiro conjunto de dados, e 42 variáveis referentes às temperaturas mínimas, médias e máximas, nos três meses mais frios e nos três meses mais quentes do ano.

Os dois conjuntos de dados foram unidos, dando origem a um único conjunto de dados formado por 12096 observações e 131 variáveis. A união dos dois conjuntos de dados foi implementada recorrendo ao cruzamento da informação relativa ao município ao qual pertence cada parcela, com as temperaturas correspondentes que se verificam nesse município. Tal permite facilitar os estudos subsequentes.

### 5.2.1 Descrição das variáveis

Tendo por base a elevada dimensionalidade inerente ao conjunto de dados, formado por 131 variáveis, foi necessário selecionar algumas variáveis, em função da sua adequação à presente investigação. Por outro lado, houve também necessidade de criar novas variáveis, nomeadamente, a variável da frequência relativa do número de árvores vivas nas parcelas, N_fr, e a classe clima.

Foi criada a variável referente à frequência relativa do número de árvores vivas, N_fr, cuja construção fez uso das variáveis 'N_vivas' e 'Npl', as quais especificam o número de árvores vivas e o número de árvores plantadas nas parcelas, respetivamente. Desta forma, um olhar crítico sob as variáveis permite intuir que, a aplicação de metodologias de gestão florestal adequadas favorece o número de árvores sobreviventes na unidade parcelar. Assim, seria interessante, como índole de qualidade da parcela, analisar a quantidade de árvores sobreviventes numa parcela e comparar com o respetivo número em outras parcelas. Não obstante, o número de árvores plantadas em cada parcela varia, e, por conseguinte, o número de árvores vivas varia também, pelo que não é correto comparar esses valores entre parcelas diferentes. No sentido de colmatar essa problemática, foi criada uma nova variável, designada por 'N_fr', que designa a frequência relativa do número de árvores que sobreviveram em cada parcela, a qual é obtida a partir da seguinte fórmula:

$$N_{fr} = \frac{N_{vivas}}{N_{plantadas}}. \tag{5.1}$$

O valor obtido a partir deste quociente é um valor relativo, pelo que a variável criada permite efetuar a comparação da quantidade de árvores sobreviventes entre parcelas, constituindo desta forma um índice de qualidade da parcela.

Numa primeira fase, tendo por base a literatura consultada (Tomé et al., 2006), considerou-se profícuo analisar a influência do clima no AMA, na presença do *Gonipterus platensis*, visto que o segundo conjunto de dados, correspondente à temperatura, só foi conseguido mais tarde. Assim, foi necessário criar uma variável referente à classe do clima por forma a facilitar as análises implementadas ao nível desta variável. A variável 'clima' que assume valores entre 1

e 10.5 foi recodificada numa nova variável, designada por 'Classe_Clima' e constituída por 5 classes, nomeadamente [1;3], [3.5; 5], [5.5; 7], [7.5; 9] e [9.5; 10.5].

Por fim, salienta-se que o critério de seleção de variáveis se baseou por um lado na influência que supostamente têm sobre a produtividade e por outro na influência que essas variáveis têm no incremento da quantidade de ataque da praga. De facto, foram selecionadas as variáveis cuja literatura consultada permite conjeturar a sua importância no ataque do *Gonipterus platensis*, bem como variáveis cuja análise exploratória implementada e conhecimentos no âmbito florestal permitiram justificar a sua importância no presente estudo. Com efeito, as variáveis usadas na análise foram:

- Acréscimo Médio Anual Útil projetado aos 12 anos (AMA_U_Proj12);
- a altitude das parcelas (cota);
- o número de dias de precipitação registados nas parcelas em estudo (dias_pp);
- a frequência do número de árvores vivas (N_fr);
- o nível de ataque do *Gonipterus platensis* (n_ataque);
- as temperaturas mínima (TMinF), média (TMedF) e máxima (TMaxF) dos três meses mais frios;
- as temperaturas mínima (TMinQ), média (TMedQ) e máxima (TMaxQ) dos três meses mais quentes;
- classe clima (Classe_Clima).

## 5.3 Análise preliminar dos dados

Nesta secção são descritas as análises que antecedem a implementação do modelo de regressão linear, em particular a análise de normalidade das variáveis envolvidas e o estudo das relações entre elas.

### 5.3.1 Estatísticas descritivas sumárias e normalidade

Fazendo-se uso do comando '*summary()*' do *software* R, são obtidas as características sumárias das variáveis em estudo (média, desvio padrão–dp, mínimo e máximo), as quais se encontram sintetizadas na tabela 5.1. É de salientar que na presente tabela as variáveis referentes ao nível de ataque e à classe clima não se encontram presentes por se tratarem de variáveis qualitativas, cada uma com 5 níveis, e, por conseguinte, nem todas as características sumárias na tabela se adequarem às mesmas. Porém quanto a variável nível de ataque é possível verificar o mínimo de 1, o máximo de 5 e a moda de 1.

Tabela 5.1: Estatística descritivas

| Variáveis | Média | dp | Min | Max |
|---|---|---|---|---|
| AMA_U_Proj12 | 8.478 | 5.305 | 0.110 | 52.918 |
| cota | 272.270 | 174.209 | 50 | 917 |
| dias_pp | 102.830 | 22.814 | 55.000 | 154.900 |
| N_fr | 0.612 | 0.204 | 0.019 | 1.000 |
| TMinF | 4.987 | 1.910 | −0.333 | 10.333 |
| TMedF | 9.538 | 1.624 | 5.333 | 13.333 |
| TMaxF | 13.607 | 1.445 | 9.333 | 17.333 |
| TMinQ | 14.872 | 1.719 | 10.667 | 29.000 |
| TMedQ | 24.436 | 1.761 | 19.000 | 32.000 |
| TMaxQ | 29.784 | 2.299 | 13.000 | 33.667 |

Relativamente aos valores sintetizados na tabela 5.1, salienta-se que, no que diz respeito à variável da frequência do número de árvores vivas, em média nas parcelas inventariadas, a quantidade de árvores que sobrevivem ronda os 60%, aproximadamente. O reduzido número verificado encontra-se intimamente relacionado com o nível de ataque do *Gonipterus platensis* na medida em que a presença de praga, impede o favorecimento do crescimento saudável das árvores existentes nas parcelas, podendo, em casos extremos, levar à morte das árvores existentes. No que se refere à variável cota, verifica-se que, em média, nas parcelas inventariadas o nível de altitude é, aproximadamente, 272 metros, altitude na qual, tal como analisado na secção 5.3.2, existe evidência da existência de praga. Tal justifica, pelo menos em parte, os valores de percentagem de árvores sobreviventes supracitados.

Tendo por base a importância inerente ao conhecimento da distribuição de probabilidade das variáveis em estudo por forma a implementar testes de Inferência Estatística, fundamentais na determinação de evidências estatísticas, começou por se obter os QQ-*plots* e os histogramas e respetivas curvas das funções densidade de probabilidade das variáveis a estudar cuja seleção se considerou profícua na presente investigação. Para tal foi implementado um código no *software* R, tendo-se feito uso dos comandos '*hist()*' e '*qqplot()*', o que permitiu obter os gráficos da figura 5.1.

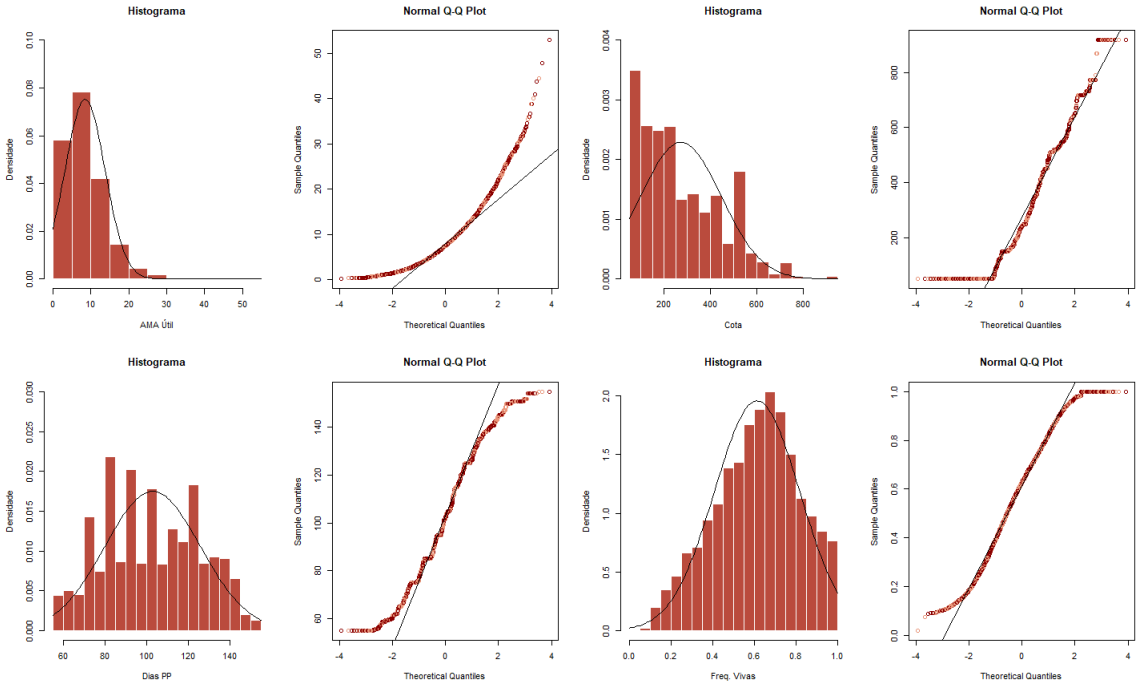



Figura 5.1: Histogramas e QQ-plots referentes às quatro variáveis

Uma análise detalhada aos histogramas apresentados permite supor a inexistência de normalidade nas variáveis selecionadas, tendo por base a inexistência de simetria nos respetivos histogramas. Também os QQ-*plots* revelam afastamento relativamente à reta que resulta da igualdade dos quantis teóricos e empíricos. No sentido de comprovar o explicado anteriormente recorreu-se a testes de normalidade que se apresentam na tabela 5.2. Com efeito, a inexistência de normalidade nas variáveis selecionadas tem como consequência a impossibilidade de aplicação de metodologias de inferência estatística cujos pressupostos incluam a existência de normalidade nos dados.

Tabela 5.2: Resultados dos testes de normalidade referentes às quatro variáveis

| | Anderson-Darling | KS com correção de Lilliefors | Decisão |
|---|---|---|---|
| AMA_U_Proj12 | $< 2.2e-16$ | $< 2.2e-16$ | Rejeita H0 |
| Cota | $< 2.2e-16$ | $< 2.2e-16$ | Rejeita H0 |
| Dias_pp | $< 2.2e-16$ | $< 2.2e-16$ | Rejeita H0 |
| N_fr | $< 2.2e-16$ | $< 2.2e-16$ | Rejeita H0 |

Nota: $p\text{-}value \leq 0.05$ Rejeita H0; $p\text{-}value > 0.05$ Não Rejeita H0

De forma análoga, foi implementado o mesmo procedimento nas variáveis referentes à temperatura. Os gráficos dos histogramas e QQ-*plots* obtidos encontram-se na figura 5.2.

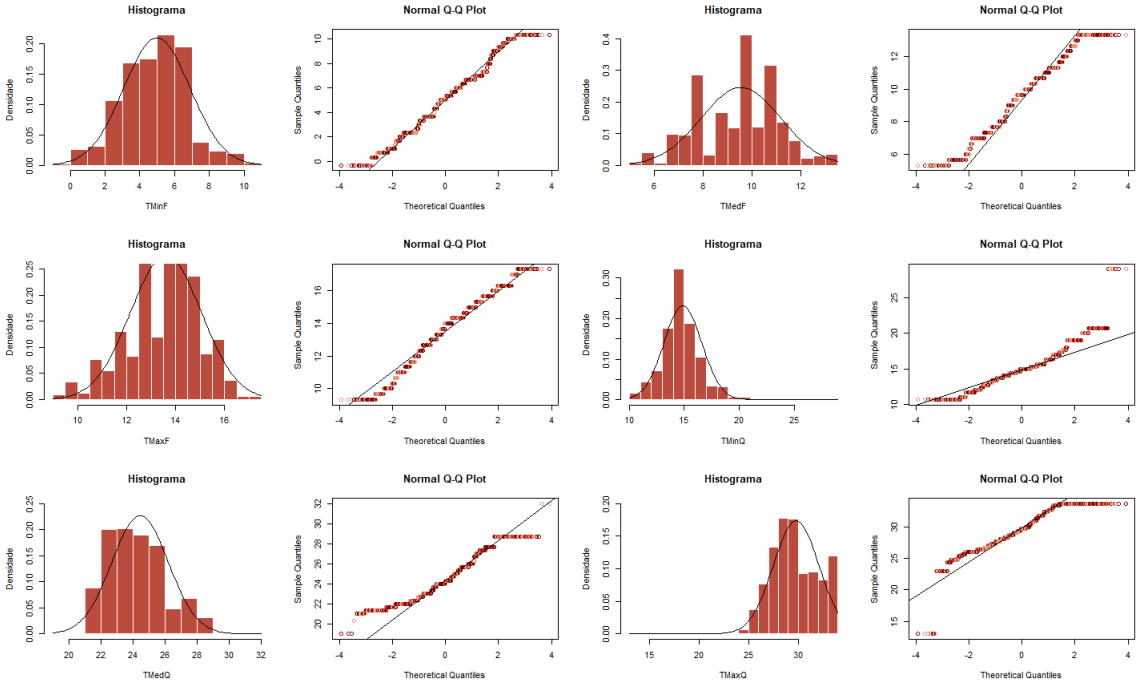



Figura 5.2: Histogramas e QQ-plots referentes às variáveis temperatura

A análise dos histogramas e respetivos QQ-*plots* apresentados na figura 5.2 mostram afastamento da distribuição normal, confirmado pelos resultados dos testes de ajustamento apresentados na tabela 5.3.

Contudo é de realçar que os testes de ajustamento (em particular o teste de KS com correção de Lilliefors) quando a dimensão é elevada têm tendência a rejeitar sempre a hipótese nula, pelo que a decisão deverá ser tomada com uma análise do QQ-*plot*. Desta forma, a aplicação de metodologias de Inferência Estatística deve ter em conta a violação de pressupostos referentes à normalidade destas variáveis.

Tabela 5.3: Resultados dos testes de normalidade referentes às variáveis temperaturas

| | Anderson-Darling | KS com correção de Lilliefors | Decisão |
|---|---|---|---|
| TMinF | $< 2.2e-16$ | $< 2.2e-16$ | Rejeita H0 |
| TMedF | $< 2.2e-16$ | $< 2.2e-16$ | Rejeita H0 |
| TMaxF | $< 2.2e-16$ | $< 2.2e-16$ | Rejeita H0 |
| TMinQ | $< 2.2e-16$ | $< 2.2e-16$ | Rejeita H0 |
| TMedQ | $< 2.2e-16$ | $< 2.2e-16$ | Rejeita H0 |
| TMaxQ | $< 2.2e-16$ | $< 2.2e-16$ | Rejeita H0 |

Nota: *p-value*$\leq$0.05 Rejeita H0; *p-value*$>$0.05 Não Rejeita H0

Nas secções subsequentes são implementadas análises exploratórias às variáveis selecionadas, as quais permitem justificar a sua utilização, bem como coadunar as análises de regressão implementadas na presente investigação.

### 5.3.2 Distribuição do ataque do *Gonipterus platensis* em Portugal continental

O mapa da figura 5.3, construído no *software* R com recurso ao *package ggmap*, utiliza mapas fornecidos pelo servidor do *Google Maps*. A sua análise e observação cuidadas permitem visualizar os níveis de ataque da praga em Portugal Continental.

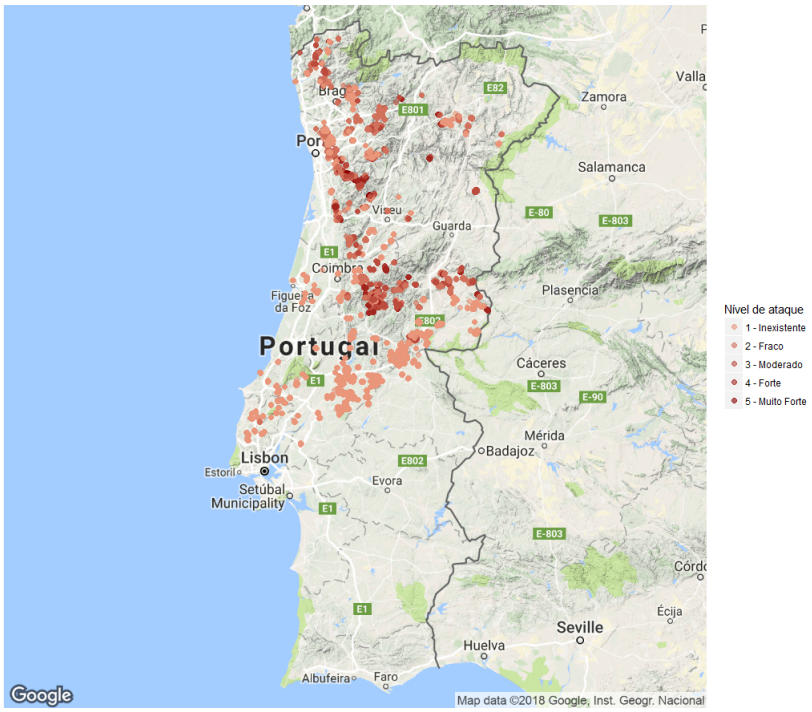



Figura 5.3: Mapa de Portugal Continental com a distribuição do ataque

Verifica-se que, de facto, o nível de ataque é variável em função da sua localização geográfica. A observação da figura 5.3 permite concluir que as parcelas inventariadas se localizam, de forma global, a norte de Lisboa. Não obstante, apenas se verifica a existência de ataque do parasita em zonas localizadas mais a norte de Portugal. Uma justificação plausível para que tal suceda, baseia-se na cota das parcelas. Assim, tendo por base a literatura consultada, o ataque é tanto maior quanto maior a cota (ou seja, a altitude) na qual se localiza a parcela, isto é, o *Gonipterus platensis* ataca de forma preferencial as parcelas que se localizam em altitudes mais elevadas.

### 5.3.3 Contabilização do número de parcelas monitorizadas

Uma análise preliminar efetuada ao conjunto de dados em estudo, baseou-se na contabilização da quantidade de parcelas por cada nível de ataque nos 5 anos em estudo. Assim, em aproximadamente metade das parcelas inventariadas verifica-se que o ataque é inexistente, isto é, o *Gonipterus platensis* apenas se verifica em, aproximadamente, metade das parcelas. Nas parcelas inventariadas onde foi verificada a presença da praga, é possível observar que apenas uma pequena minoria apresenta evidências de ataque muito forte. Com efeito, na maioria das parcelas inventariadas que são afetadas pela praga verifica-se que o ataque é fraco ou moderado.

Relativamente às parcelas inventariadas onde predominavam evidencias da existência de níveis de ataque muito fortes, pode-se dizer que o seu uso para a entidade de acolhimento,

RAIZ, é substancialmente reduzido na medida em que a madeira aproveitada após um ataque tão forte do *Gonipterus platensis* se pauta por quantidades diminutas. Com efeito, em parcelas onde este nível de ataque se verifique, o aproveitamento de madeira e outras matérias primas é praticamente nulo.

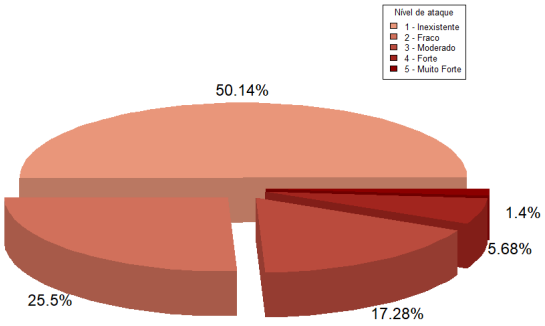



Figura 5.4: Percentagem de parcelas por nível de ataque

Com o objetivo de analisar, detalhadamente, o nível de ataque ao longo dos anos, foram construídos gráficos de barras, tendo-se feito uso para tal da função *barplot*() do *software* R. Uma análise do gráfico de barras da figura 5.5 sugere um aumento da quantidade de ataques desde 2011 até 2015, visível pelo aumento do tamanho das barras ao longo dos anos em estudo. Não obstante, verifica-se que o aumento do número de parcelas avaliadas por nível de ataque também aumenta ao longo dos anos, tornando-se difícil concluir sobre o aumento efetivo de quantidade de ataques.

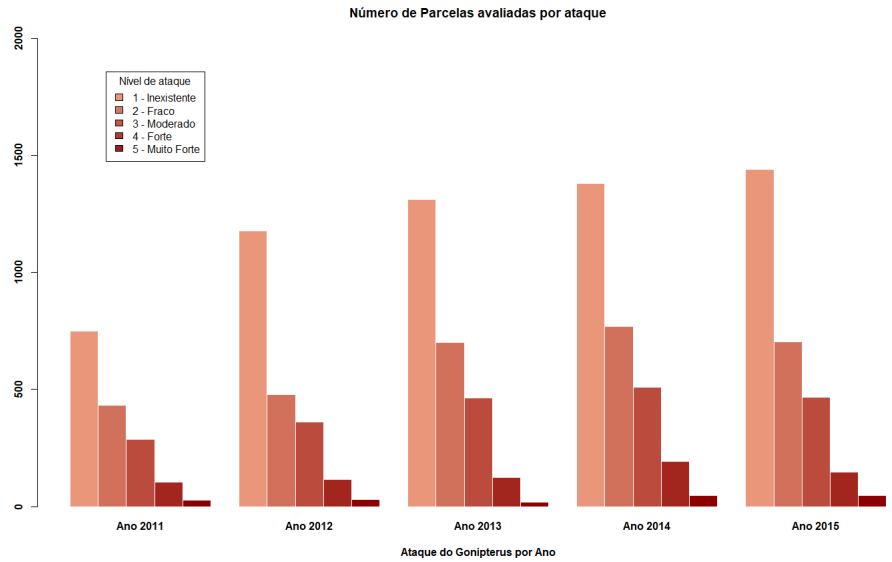



Figura 5.5: Número de parcelas por nível de ataque

Assim, foram obtidos os respetivos valores percentuais, para desta forma, comprovar ou refutar um aumento da quantidade de ataque verificado ao longo dos anos em estudo. Os resultados obtidos relativos aos valores percentuais do nível de ataque por ano, sintetizam-se na tabela 5.4.

Tabela 5.4: Nível de ataque por ano em percentagem

| | Ano 2011 | Ano 2012 | Ano 2013 | Ano 2014 | Ano 2015 |
|---|---|---|---|---|---|
| 1 - Inexistente | 46.94 | 54.46 | 50.04 | 47.59 | 51.34 |
| 2 - Fraco | 27 | 22.12 | 26.75 | 26.48 | 25.06 |
| 3 - Moderado | 17.88 | 16.72 | 17.72 | 17.60 | 16.68 |
| 4 - Forte | 6.56 | 5.36 | 4.76 | 6.68 | 5.24 |
| 5 - Muito Forte | 1.62 | 1.34 | 0.72 | 1.65 | 1.68 |

A análise dos valores percentuais do nível de ataque em cada ano em estudo permite concluir que ao longo dos 5 anos, aproximadamente metade das parcelas não apresentam evidências da existência de praga.

Uma análise percentual da quantidade de parcelas que, em cada ano em estudo, apresentavam os diferentes níveis de ataque, permite verificar que ao longo dos anos não há um aumento significativo da quantidade de parcelas em cada nível ataque. Para cada nível de ataque, verifica-se que a percentagem de parcelas em cada ano se mantém praticamente constante, ou que apresenta valores bastante semelhantes.

Com efeito, uma justificação plausível para a existência de um aumento do tamanho das barras do gráfico da figura 5.5, e, em contrapartida, a existência de valores percentuais semelhantes da quantidade de ataque por ano, baseia-se num incremento do número de parcelas inventariadas e não num aumento da quantidade de ataque existente ao longo dos anos.

Salienta-se que a inexistência de praga foi mais frequente no ano 2012, e, por outro lado, verifica-se que no ano de 2013 a quantidade de parcelas cujo ataque do *Gonipterus platensis* foi classificado como sendo muito forte, foi mais reduzido, de entre as restantes parcelas classificadas por este nível de ataque, nos restantes anos em estudo.

As parcelas cujo nível de ataque é classificado como sendo muito forte ou forte, em termos percentuais, são as que apresentam valores mais reduzidos, sendo o valor máximo na ordem dos 7%, aproximadamente. No que diz respeito às parcelas classificadas pelos níveis de ataque fraco e moderado, verifica-se que as diferenças percentuais entre os mesmos são bastante reduzidas, assumindo valores percentuais bastante semelhantes entre si, pelo que estes níveis de ataque se apresentam como idênticos ao longo dos anos em estudo. Uma justificação plausível para tal

suceder, pode ser a existência de monitorização e tratamento da praga em parcelas classificadas por um dos níveis de ataque – fraco ou moderado – o que dificulta e minimiza a propagação da praga, minimizando as consequências nefastas inerentes à sua atuação. Não obstante, na presente investigação, a falta de dados relativos à monitorização e tratamento da praga, impossibilitam comprovar a afirmação supracitada.

### 5.3.4 Análise da variável AMA Útil Proj12

Tendo por base conhecimentos no âmbito florestal e o levantamento bibliográfico desenvolvido, foi possível tomar consciência, a nível teórico, das consequências nefastas inerentes à presença do *Gonipterus platensis* em áreas de minifúndio florestal. Com efeito, a presença da praga, ou gorgulho do eucalipto, em áreas florestais, minimiza a quantidade de matéria prima extraída, e, em certos casos extremos e de elevada gravidade pode levar mesmo à perda total da madeira existente nessas áreas. Assim, seguidamente objetiva-se tomar consciência da influência da presença do gorgulho do eucalipto nas zonas em estudo, na produtividade das parcelas. Para tal, foram desenvolvidas inúmeras análises nesse sentido, as quais relacionam o nível de ataque com o AMA_U_Proj12. A escolha desta variável prende-se com a necessidade de selecionar uma variável que permita avaliar a produtividade das parcelas. Assim, sendo esta variável pautada de elevada importância para o estudo da produtividade da parcela, e sendo também uma variável de interesse para a entidade acolhedora, os estudos desenvolvidos basearam-se nela. O AMA_U_Proj12 permite avaliar o acréscimo médio anual, em $m^3$/ha/ano, o que traduz a quantidade de volume que, em cada ano e por cada hectare, a parcela permite incrementar, em termos de matéria prima útil.

Assim, a análise preliminar inicia-se pela representação gráfica de caixas de bigodes para os diferentes níveis de ataque e o AMA_U_Proj12 das respetivas parcelas – figura 5.6.

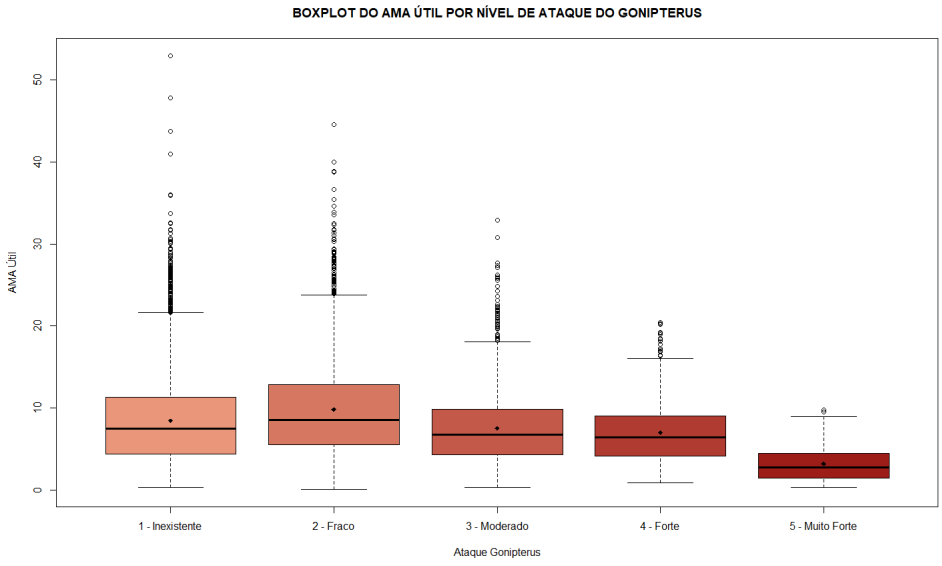



Figura 5.6: Caixa de bigodes do AMA Médio por nível de ataque

As caixas de bigodes apresentadas na figura 5.6 permitem observar que, tal como seria expectável tendo por base o levantamento bibliográfico desenvolvido, o AMA_U_Proj12 é mais reduzido em parcelas cujo nível de ataque seja classificado como forte ou muito forte, comparativamente com os restantes níveis de ataque. Tal é visível pela localização inferior das duas caixas de bigodes referentes aos dois níveis de ataque supracitados, comparativamente aos restantes. Com efeito, a análise gráfica das caixas de bigodes da figura 5.6 permite evidenciar que o valor médio do AMA_U_Proj12 (indicado como um ponto no interior das caixas de bigodes) é mais reduzido em parcelas cujo ataque é classificado como muito forte e forte, sendo mais elevado em parcelas cujo ataque é inexistente ou fraco, seguido pelas parcelas cujo nível de ataque é classificado como moderado. A observação das caixas de bigodes leva a supor que, à medida que o nível de ataque das parcelas aumenta, o seu AMA_U_Proj12, e, por conseguinte, a produtividade das parcelas diminui.

Sendo que graficamente parecem existir diferenças significativas entre os níveis de ataque, recorreu-se ao teste de Kruskal-Wallis, por forma a verificar essa suposição, em que resultou na obtenção de um *p-value* $< 2.2e-16$, sendo muito inferior que o nível de significância de 0.05. Pode, por conseguinte, afirmar-se a existência de diferenças significativas entre os grupos de tratamento. Posto isto, procedeu-se à análise *post-hoc* o que permitiu concluir que existem diferenças significativas entre todos os níveis de ataque, excetuando entre os níveis de ataque moderado e forte, o que também é possível verificar na caixa de bigodes anterior. Os resultados dos *p-values* obtidos através da análise *post-hoc* encontram-se na tabela 5.5.

Tabela 5.5: Respetivos *p-values* da análise *post-hoc*

| | 1 – Inexistente | 2 – Fraco | 3 – Moderado | 4 – Forte |
|---|---|---|---|---|
| 2 – Fraco | $2.1e-14$ | — | — | — |
| 3 – Moderado | $1.8e-08$ | $< 2e-16$ | — | — |
| 4 – Forte | $3.8e-08$ | $< 2e-16$ | 0.3 | — |
| 5 – Muito Forte | $< 2e-16$ | $< 2e-16$ | $< 2e-16$ | $1.4e-14$ |

Nota: *p-value*≤0.05 Rejeita H0; *p-value*>0.05 Não Rejeita H0

A figura 5.7 permite evidenciar comportamentos semelhantes no que se refere à função densidade de probabilidade da variável AMA_U_Proj12 em função do nível de ataque. Com efeito, é possível notar a existência de um comportamento semelhante do AMA_U_Proj12 nos níveis de ataque fraco e inexistente. No entanto, os níveis de ataques referidos distinguem-se dos restantes, em particular, dos níveis de ataque forte e muito forte, os quais possuem comportamentos distintos dos primeiros. De facto, nos níveis de ataque forte e muito forte é possível destacar a existência de picos, mais salientes no nível de ataque muito forte, para valores de AMA_U_Proj12 reduzidos, o que mostra a existência de uma elevada quantidade de parcelas cujo AMA_U_Proj12 é mais reduzido.

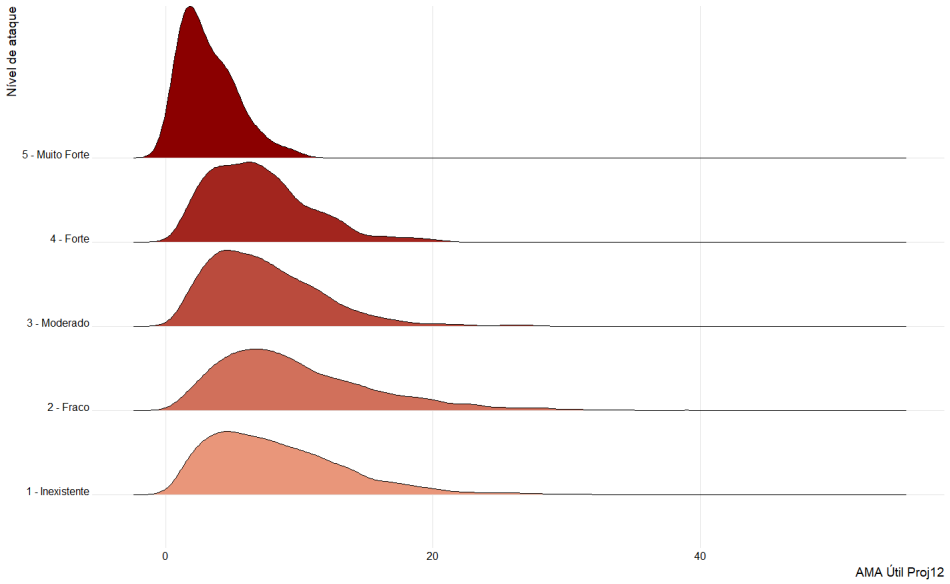



Figura 5.7: Gráfico da função densidade de probabilidade do AMA_U_Proj12 por nível de ataque

A tabela 5.6 apresenta as seguintes características sumárias, frequência absoluta (contagem), média, desvio padrão (dp), mediana e amplitude do intervalo interquartil (AIQ) da variável AMA_U_Proj12, para cada nível de ataque.

Tabela 5.6: Estatísticas sumárias do AMA_U_Proj12 por nível de ataque

| grupo | contagem | média | dp | mediana | AIQ |
|---|---|---|---|---|---|
| 1 – Inexistente | 6065 | 8.440280 | 5.296829 | 7.494092 | 6.898122 |
| 2 – Fraco | 3085 | 9.830075 | 5.882157 | 8.558873 | 7.345439 |
| 3 – Moderado | 2090 | 7.501755 | 4.340124 | 6.746769 | 5.542949 |
| 4 – Forte | 687 | 7.007965 | 3.735712 | 6.468991 | 4.866611 |
| 5 – Muito Forte | 169 | 3.186243 | 2.052884 | 2.777356 | 3.013575 |

É interessante notar que tanto o valor médio como o valor mediano do AMA_U_Proj12 é superior para o nível de ataque fraco relativamente ao nível de ataque inexistente, o que à partida, contraria o que se esperava observar. No entanto, o número de observações em cada grupo não é comparável e esta observação pode ser uma mera consequência da elevada variabilidade da classe ataque inexistente.

Atendendo a que as temperaturas assumem elevada importância na presente investigação, seguidamente são implementadas análises exploratórias às variáveis que fazem referência às temperaturas das parcelas. Foi usada informação relativa à média das temperaturas mínimas, médias e máximas dos três meses mais frios e dos três meses mais quentes.

A tabela seguinte permite tomar conhecimento dos valores médios de cada uma das variáveis referentes às temperaturas nas parcelas.

Tabela 5.7: Valores médios das variáveis temperatura

| | Temperatura média |
|---|---|
| TMinF | 4.987 |
| TMedF | 9.538 |
| TMaxF | 13.607 |
| TMinQ | 14.872 |
| TMedQ | 24.436 |
| TMaxQ | 29.784 |

Seguidamente serão implementadas diversas análises que permitem relacionar as variáveis referentes às temperaturas das parcelas, com a variável AMA_U_Proj12.

A figura 5.8 relaciona as variáveis referentes às temperaturas nas parcelas com o AMA_U_Proj12. O estudo da relação existente entre as variáveis temperaturas das parcelas com o AMA_U_Proj12 coaduna-se na medida em que, conhecimentos na área florestal permitem intuir quanto à relação existente entre as variáveis supracitadas. No que diz respeito às variáveis referentes às temperaturas nos meses mais frios, verifica-se que, à medida que as temperaturas nestes meses aumentam, o AMA_U_Proj12 aumenta. Não obstante, no âmbito das temperaturas nos meses mais quentes, verifica-se que à medida que essas temperaturas aumentam, os valores do AMA_U_Proj12 têm tendência a diminuir.

Para o efeito do seguinte gráfico considera-se o AMA_U_Proj12 Médio em função da temperatura, o que se deve ao facto de que existem vários municípios, e por conseguinte, várias parcelas em que as medições de temperatura são exatamente as mesmas, pois as estações meteorológicas não se encontram em cada parcela avaliada. Se não fosse considerado o AMA_U_Proj12 Médio o gráfico seguinte teria vários pontos de AMA_U_Proj12 para a mesma temperatura. Para isso não acontecer, para cada variável de temperatura (TMinF, TMedF, TMaxF, TMinQ, TMedQ e TMaxQ) foi calculado o AMA_U_Proj12 médio, agregando todas as parcelas correspondentes ao mesmo município.

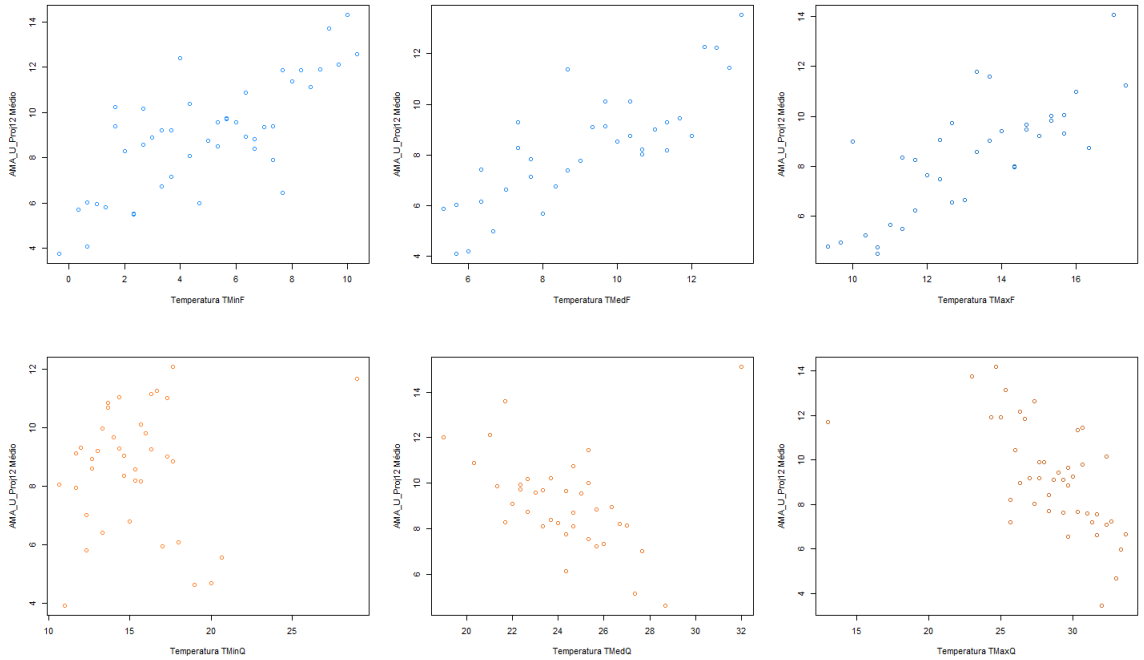



Figura 5.8: Gráficos de dispersão do AMA_U_Proj12 Médio em função da temperatura

Assim, a figura 5.8 permite evidenciar que, de facto, deverá haver uma relação entre a temperatura e o AMA_U_Proj12 médio. Com efeito, de uma forma geral o AMA_U_Proj12 médio aumenta à medida que a temperatura referentes aos três meses mais frios aumenta. Porém, quando as temperaturas dizem respeito às variáveis dos três meses mais quentes é possível verificar uma diminuição na variável AMA_U_Proj12 médio quando o valor das temperaturas é superior a aproximadamente 18ºC.

De maneira a compreender, de forma mais detalhada, a relação existente entre as temperaturas, o nível de ataque e o AMA_U_Proj12 das parcelas, foram construídos os modelos de regressão linear da variável AMA_U_Proj12 considerando separadamente cada uma das variáveis temperatura em função de cada um dos níveis de ataque. Os respetivos gráficos são apresentados na figura 5.9.

O modelo genérico estimado referente aos modelos de regressão linear construídos é o seguinte:

$$\begin{aligned} \widehat{AMA\_U\_Proj12} = \hat{\beta}_0 + \hat{\beta}_1 n\_ataque1 + \hat{\beta}_2 n\_ataque2 + \hat{\beta}_3 n\_ataque3 \\ + \hat{\beta}_4 n\_ataque4 + \hat{\beta}_5 n\_ataque5 + \hat{\beta}_6 temp \end{aligned} \tag{5.2}$$

onde *temp* pode representar cada uma das seguintes variáveis TMinF, TMedF, TMaxF, TMinQ, TMedQ e TMaxQ.

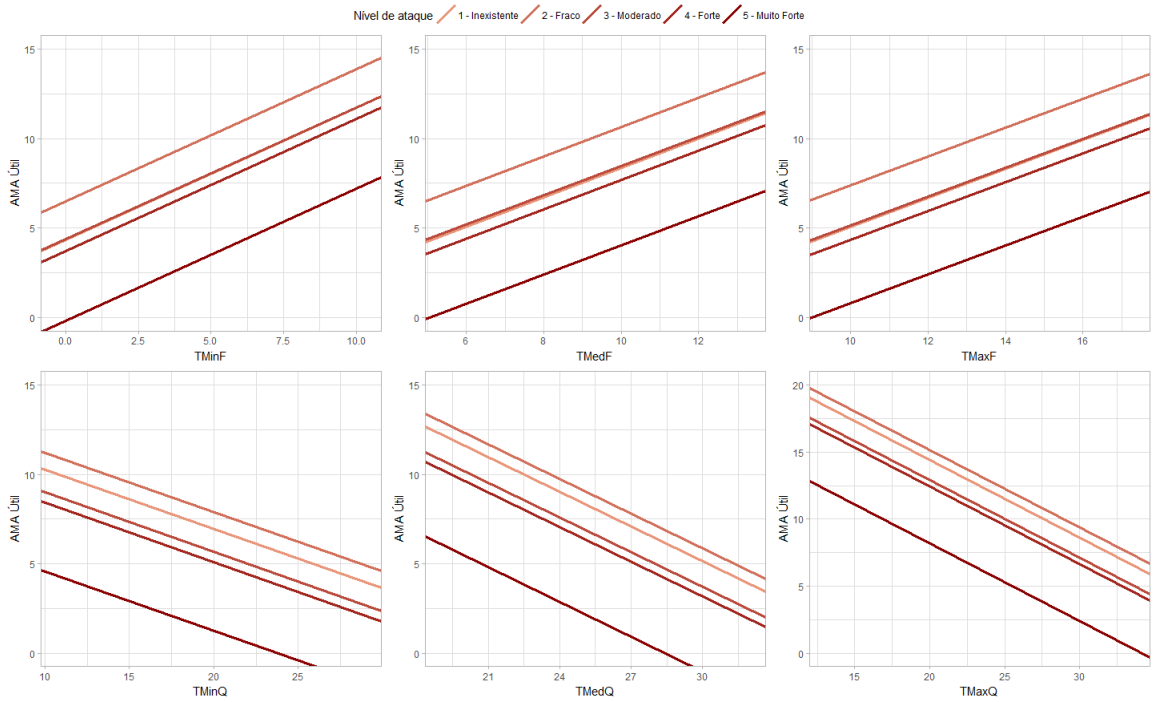



Figura 5.9: Gráficos do AMA_U_Proj12 em função da temperatura

Os gráficos da figura 5.9 permitem evidenciar, uma vez mais, que, no que diz respeito às temperaturas nos meses mais frios, à medida que as temperaturas aumentam, os valores de AMA_U_Proj12 aumentam também. Os resultados dos modelos de regresssão OLS AMA_U_Proj12 relativos às temperaturas acima referidas em função de cada um dos níveis de ataque, estão apresentados na tabela 5.8. No corpo central da tabela encontram-se os valores das estimativas dos parâmetros de cada um dos 6 modelos de regressão considerados, numerados de (1) a (6), e os respetivos erros padrão (entre parêntesis).

Tabela 5.8: Resultados dos modelos de regressão OLS AMA_U_Proj12 relativamente a cada uma das variáveis temperatura em função de cada um dos níveis de ataque

| | *Variável dependente:* | | | | | |
|---|---|---|---|---|---|---|
| | AMA_U_Proj12 | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Constant | 4.353*** | 0.193 | −2.977*** | 13.571*** | 24.530*** | 25.962*** |
| | (0.154) | (0.312) | (0.499) | (0.456) | (0.681) | (0.623) |
| n_ataque 2 | 2.169*** | 2.269*** | 2.298*** | 0.961*** | 0.724*** | 0.746*** |
| | (0.114) | (0.116) | (0.119) | (0.120) | (0.116) | (0.114) |
| n_ataque 3 | 0.026 | 0.111 | 0.068 | −1.268*** | −1.434*** | −1.467*** |
| | (0.132) | (0.134) | (0.136) | (0.134) | (0.131) | (0.129) |
| n_ataque 4 | −0.638*** | −0.666*** | −0.725*** | −1.854*** | −1.984*** | −1.965*** |
| | (0.204) | (0.205) | (0.207) | (0.211) | (0.206) | (0.203) |
| n_ataque 5 | −4.516*** | −4.319*** | −4.277*** | −5.689*** | −6.164*** | −6.236*** |
| | (0.392) | (0.395) | (0.399) | (0.405) | (0.398) | (0.394) |
| TMinF | 0.735*** | — | — | — | — | — |
| | (0.025) | | | | | |
| TMedF | — | 0.816*** | — | — | — | — |
| | | (0.030) | | | | |
| TMaxF | — | — | 0.805*** | — | — | — |
| | | | (0.035) | | | |
| TMinQ | — | — | — | −0.332*** | — | — |
| | | | | (0.029) | | |
| TMedQ | — | — | — | — | −0.646*** | — |
| | | | | | (0.027) | |
| TMaxQ | — | — | — | — | — | −0.578*** |
| | | | | | | (0.020) |
| Observations | 12,096 | 12,096 | 12,096 | 12,096 | 12,096 | 12,096 |
| $R^2$ | 0.104 | 0.095 | 0.081 | 0.051 | 0.083 | 0.100 |
| Adjusted $R^2$ | 0.104 | 0.095 | 0.081 | 0.050 | 0.083 | 0.100 |
| Residual Std. Error | 5.021 | 5.046 | 5.086 | 5.169 | 5.080 | 5.033 |
| F Statistic (df = 5; 12090) | 281.487*** | 254.961*** | 213.611*** | 129.559*** | 219.972*** | 269.156*** |

Nota: *p<0.1; **p<0.05; ***p<0.01

Uma análise à tabela anterior, nomeadamente aos valores do teste F de significância global dos modelos de regressão linear e o respetivo *p-value*, permite rejeitar a hipótese nula dos coeficientes serem todos iguais a zero, logo, o modelo de regressão de AMA_U_Proj12 em

função de cada uma das variáveis da tabela é significativo como um todo, para qualquer um dos 6 modelos ajustados relativamente a cada uma das variáveis temperatura.

Seguidamente foram analisadas as relações existentes entre o valor do AMA Útil projetado a 12 anos e algumas variáveis consideradas preponderantes na presente investigação. As variáveis em estudo são: a altitude, os dias de precipitação e a frequência relativa do número de árvores vivas.

Por forma a prosseguir com essa análise recorreu-se a gráficos de dispersão que relacionam o AMA_U_Proj12 com as três variáveis já referidas.

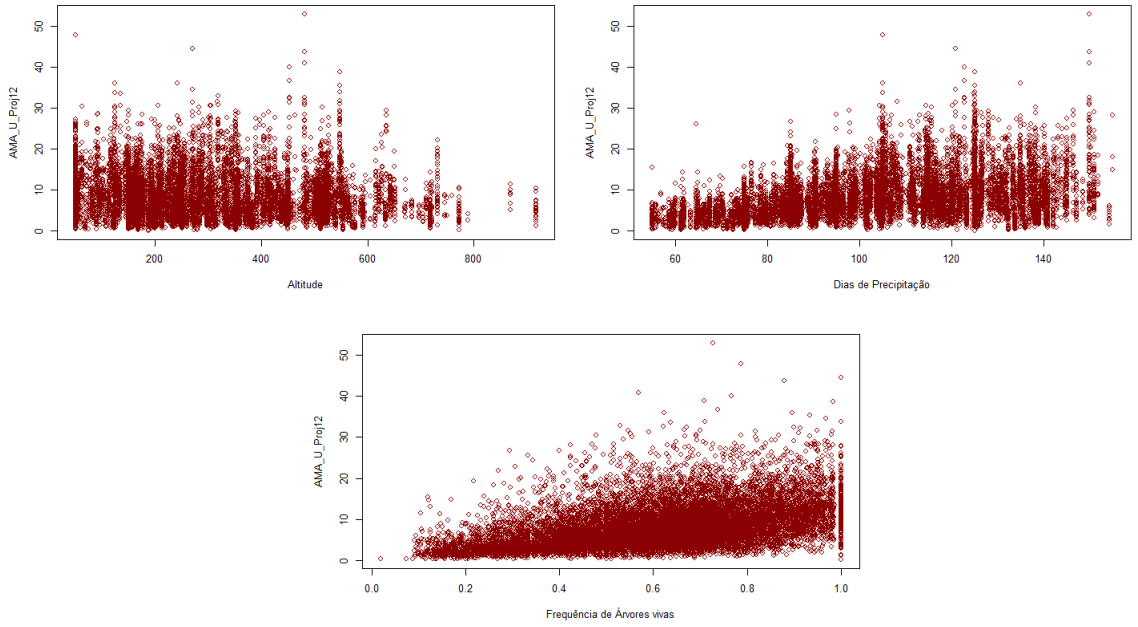



Figura 5.10: Gráficos de dispersão do AMA_U_Proj12 em função das 3 variáveis

Pela figura 5.10 é possível verificar que relativamente ao gráfico que diz respeito à altitude, de uma forma geral, o AMA_U_Proj12 diminui à medida que a altitude aumenta, em particular a partir de um nível de altitude de cerca de 600m. Contrariamente ao que acontecia no primeiro gráfico, no que se refere aos outros dois gráficos de dispersão, o AMA_U_Proj12 aumenta à medida que tanto os dias de precipitação como a frequência de árvores vivas aumentam.

Portanto, de forma a ir ao encontro do que foi explicado anteriormente, procedeu-se a uma análise mais detalhada, implementando-se um modelo de regressão linear da variável AMA_U_Proj12 relativamente a cada uma das três variáveis em função do nível de ataque existente na parcela.

Os modelos de regressão linear construídos são traduzidos analiticamente pela expressão:

$$\widehat{AMA\_U\_Proj12} = \hat{\beta}_0 + \hat{\beta}_1 n\_ataque1 + \hat{\beta}_2 n\_ataque2 + \hat{\beta}_3 n\_ataque3 + \hat{\beta}_4 n\_ataque4 + \hat{\beta}_5 n\_ataque5 + \hat{\beta}_6 var \tag{5.3}$$

onde *var* representa cada uma das variáveis independentes altitude, dias de precipitação e frequência do número de árvores vivas.

Segue-se a representação gráfico dos referidos modelos.

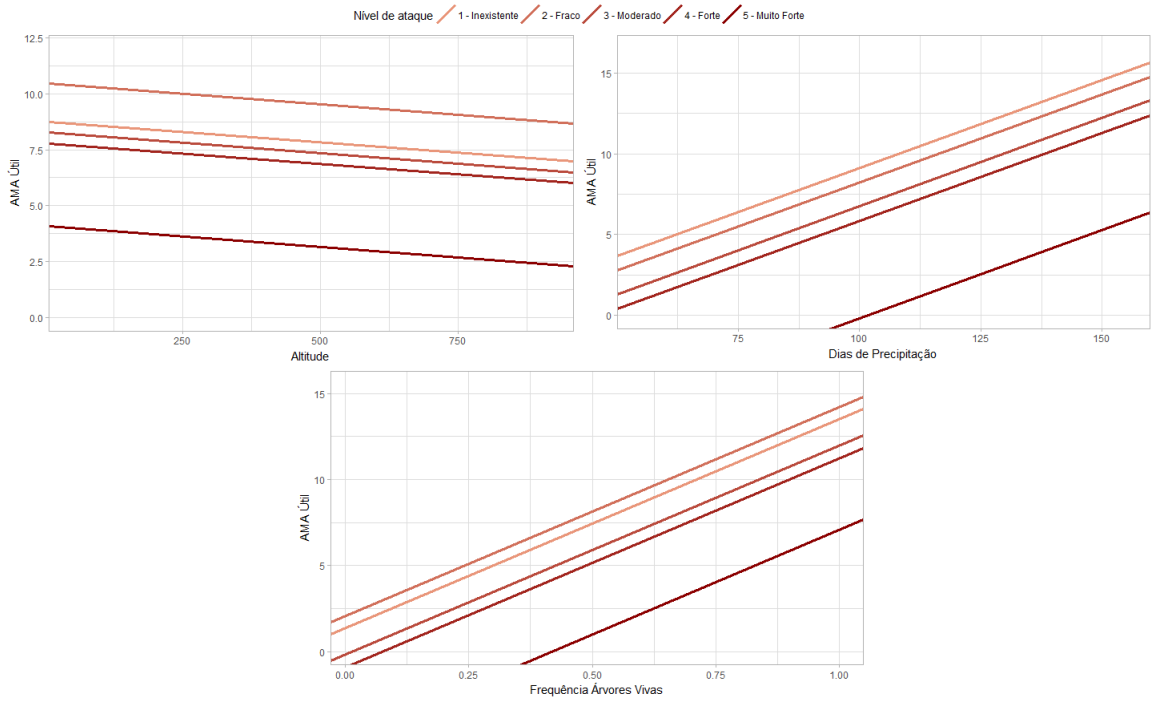



Figura 5.11: Gráficos do AMA_U_Proj12 em função das 3 variáveis

Os gráficos apresentados permitem intuir que, tal como explanado na secção 5.3.2, à medida que a altitude (cota) das parcelas aumenta, o valor do AMA_U_Proj12 diminui. Relativamente às variáveis número de dias de precipitação e frequência relativa do número de árvores vivas, à medida que cada uma delas aumenta, os valores do AMA_U_Proj12 aumentam de forma linear. Verifica-se que as parcelas pertencentes aos diferentes níveis de ataque possuem o mesmo comportamento para as três variáveis independentes, sendo que parcelas cujo nível de ataque é mais forte possuem valores e AMA_U_Proj12 mais reduzido do que as parcelas classificadas com os restantes níveis de ataque.

Os resultados dos modelos de regressão OLS AMA_U_Proj12 relativos às variáveis acima referidas em função de cada um dos níveis de ataque, encontram-se sintetizados na tabela 5.9.

Tabela 5.9: Resultados dos modelos de regressão OLS AMA_U_Proj12 considerando cada um dos modelos em função dos níveis de ataque, onde (1), (2) e (3) designam os modelos de regressão considerando a variável explicativa cota, dias_pp e N_fr, respetivamente.

| | *Variável dependente:* | | |
|---|---|---|---|
| | AMA_U_Proj12 | | |
| | (1) | (2) | (3) |
| Constant | 8.765*** | −1.785*** | 1.361*** |
| | (0.089) | (0.203) | (0.134) |
| n_ataque 2 | 1.685*** | −0.902*** | 0.700*** |
| | (0.127) | (0.112) | (0.102) |
| n_ataque 3 | −0.499*** | −2.375*** | −1.535*** |
| | (0.154) | (0.122) | (0.117) |
| n_ataque 4 | −0.967*** | −3.283*** | −2.285*** |
| | (0.225) | (0.192) | (0.185) |
| n_ataque 5 | −4.678*** | −9.310*** | −6.439*** |
| | (0.418) | (0.373) | (0.358) |
| cota | −0.002*** | — | — |
| | (0.0003) | | |
| dias_pp | — | 0.109*** | — |
| | | (0.002) | |
| N_fr | — | — | 12.134*** |
| | | | (0.207) |
| Observations | 12,096 | 12,096 | 12,096 |
| $R^2$ | 0.043 | 0.221 | 0.254 |
| Adjusted $R^2$ | 0.043 | 0.220 | 0.253 |
| Residual Std. Error (df = 12090) | 5.190 | 4.684 | 4.583 |
| F Statistic (df = 5; 12090) | 109.038*** | 684.890*** | 821.996*** |
| Nota: | | | *p<0.1; **p<0.05; ***p<0.01 |

No que refere aos valores do teste F de significância global dos modelos de regressão linear e o respetivo *p-value*, pode assim concluir-se a rejeição da hipótese nula dos coeficientes serem todos iguais a zero. Desta forma, o modelo de regressão de AMA_U_Proj12 em função de cada uma das variáveis apresentadas na tabela é significativo como um todo.

## 5.4 Modelo de regressão linear múltipla

Com o intuito de verificar e inferir quanto à existência de fatores que influenciam o Acréscimo Médio Anual projetado aos 12 anos (AMA_U_Proj12) das parcelas, bem como o sentido em que essa influência ocorre e a sua dimensão, efetua-se em seguida uma análise de regressão que inclui todas as variáveis consideradas pertinentes nas análises que se descreveram na secção anterior. O principal objetivo consiste na construção de um modelo capaz de estimar a produtividade quando se verificam diferentes níveis de ataque do *Gonipterus platensis* nas parcelas.

Note-se que esta análise é realizada sem as variáveis TMedF e TMedQ por se saber, à partida, que se tratam de variáveis que resultam da combinação linear das restantes variáveis temperatura.

Tendo sido selecionadas as variáveis mencionadas nas secção anterior, começou-se por calcular os coeficientes de correlação entre cada uma dessas variáveis e o AMA_U_Proj12. Os resultados apresentam-se na tabela 5.10.

Tabela 5.10: Correlações entre as variáveis AMA_U_Proj12 com as restantes

| | AMA_U_Proj12 |
|---|---|
| TMinF | 0.245 |
| TMaxF | 0.185 |
| TMinQ | −0.104 |
| TMaxQ | −0.241 |
| Cota | −0.071 |
| Ataque | −0.095 |
| N_fr | 0.473 |
| dias_pp | 0.394 |

Inicialmente foram implementados dois modelos de regressão linear, nos quais se considerou como variável dependente o AMA_U_Proj12. Para cada modelo foram construídos subconjuntos das variáveis independentes supracitadas, o que permitiu verificar e selecionar o melhor modelo de regressão cujas variáveis explicativas mais influenciavam o AMA_U_Proj12 na presença do ataque do parasita.

A construção dos modelos supracitados fez uso da função *lm()* do *software* R, que intrinsecamente aplica o método dos mínimos quadrados ordinários.

Tal como explanado anteriormente foram construídos dois modelos de regressão linear múltipla, onde se considera o AMA_U_Proj12 como variável dependente, , sendo que o modelo 1 (descrito no anexo C.1) contém todas as variáveis mencionadas na tabela 5.10.

A justificação da implementação de um modelo distinto do modelo 1 reside na necessidade de colmatar a verificação de que a inexistência de ataque não influenciava o AMA_U_Proj12. De facto, não seria expectável que tal sucedesse, uma vez que se espera que a inexistência de ataque nas parcelas influencie o seu valor de AMA_U_Proj12.

Analogamente ao modelo 1, o presente modelo foi construído utilizando o AMA_U_Proj12 como variável dependente, diferindo do primeiro em relação às variáveis independentes constituintes do modelo. Assim, foram usadas todas as variáveis do primeiro modelo, tendo sido apenas alteradas as variáveis referentes à temperatura. Tendo-se observado que a aplicação da metodologia *Stepwise* ao modelo 1 permitiu retirar a variável TMaxF, restaram as variáveis TMinQ, TMaxQ e TMinF. A partir da observação da tabela 5.10 a seleção final recaiu nas variáveis da temperatura mínima dos meses mais frios e da temperatura máxima dos meses mais quentes, visto que são as temperaturas que mais influenciam o AMA_U_Proj12.

A construção do modelo 2 permitiu alcançar os resultados apresentados na tabela 5.11.

Tabela 5.11: Resultados da aplicação da regressão OLS ao modelo 2

| | *Variável dependente:* |
|---|---|
| | AMA_U_Proj12 |
| n_ataque 2 | $-0.876^{***}$ |
| | (0.117) |
| n_ataque 3 | $-2.526^{***}$ |
| | (0.132) |
| n_ataque 4 | $-3.562^{***}$ |
| | (0.189) |
| n_ataque 5 | $-9.345^{***}$ |
| | (0.351) |
| TMinF | $0.143^{***}$ |
| | (0.024) |
| TMaxQ | $-0.129^{***}$ |
| | (0.019) |
| cota | $0.001^{**}$ |
| | (0.0003) |
| N_fr | $9.966^{***}$ |
| | (0.197) |
| dias_pp | $0.076^{***}$ |
| | (0.002) |
| Constant | $-1.501^{**}$ |
| | (0.716) |
| Observations | 12,096 |
| $R^2$ | 0.366 |
| Adjusted $R^2$ | 0.365 |
| Residual Std. Error | 4.227 (df = 12086) |
| F Statistic | $773.733^{***}$ (df = 9; 12086) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

A aplicação da metodologia de seleção de variáveis *Stepwise* não produziu resultados vantajosos neste modelo, visto que não removeu nenhuma das variáveis consideradas no modelo base. (Observável na tabela 5.12)

Tabela 5.12: Resultados do *Stepwise* no modelo 2

| | Df | Sum of Sq | RSS | AIC |
|---|---|---|---|---|
| <none> | | | 215921 | 34881 |
| − cota | 1 | 111 | 216032 | 34885 |
| − TMinF | 1 | 622 | 216543 | 34914 |
| − TMaxQ | 1 | 789 | 216710 | 34923 |
| − n_ataque | 4 | 19001 | 234922 | 35893 |
| − dias_pp | 1 | 19944 | 235865 | 35948 |
| − N_fr | 1 | 45702 | 261623 | 37201 |

*Note:* (+) Remoção da variável; (−) Manter variável

De acordo com o teste de significância de cada coeficiente, os valores do $p - value$ são significativos, a um nível de significância de 5%, visto que todos os valores são inferiores a 0.05. Não obstante, o coeficiente relativo à variável cota é ligeiramente superior aos restantes, continuando, apesar disso, a ser significativo. O mesmo ocorre ao nível do $p-value$ do teste de significância do coeficiente de referência (inexistência de ataque). Com efeito, de uma forma geral, os coeficientes inerentes às variáveis consideradas são significativamente diferentes de zero.

Quanto ao ajuste dos modelos de regressão linear, verifica-se que o valor do $R^2$ obtido foi de 0.3651, porém deveria ser próximo de um (apontando assim para uma maior proporção de variabilidade explicada pelo modelo de regressão linear relativamente ao que respeita à variável dependente). Os valores reduzidos de $R^2$ alcançados justificam-se na medida em que o modelo de regressão construído usa variáveis explicativas de índole quantitativa e qualitativa, cuja aplicação, de forma geral, implica o declínio do $R^2$.

No que concerne ao valor da estatística de significância global do modelo de regressão linear é plausível verificar que esse valor foi da ordem 773.733, cujo valor é considerado significativo. Sendo este superior a 1, é plausível concluir que o presente modelo se adequa aos dados, isto é, face à dimensionalidade dos dados o valor obtido induz a capacidade das variáveis preditoras explicarem a variável dependente.

Um dos pressupostos do modelo de regressão linear é a inexistência de colinearidade entre as variáveis dependentes. Assim, numa primeira fase da análise da adequação do modelo de regressão linear construído, objetiva-se validar este pressuposto. Para tal, fez-se uso da função '*vif()*' do *software* R. O valor do VIF (*Variance Inflation Factor*) indica o grau de colinearidade entre as variáveis independentes. A aplicação da função permitiu obtenção dos

resultados sintetizados na tabela C.4. Os resultados obtidos permitem concluir a ausência de colinearidade nas variáveis independentes.

Tabela 5.13: Resultados da aplicação do *vif()* no modelo 2

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| n_ataque | 2.065 | 4 | 1.095 |
| TMinF | 1.457 | 1 | 1.207 |
| TMaxQ | 1.346 | 1 | 1.160 |
| cota | 1.710 | 1 | 1.308 |
| N_fr | 1.092 | 1 | 1.045 |
| dias_pp | 1.818 | 1 | 1.348 |

### 5.4.1 Análise dos resíduos

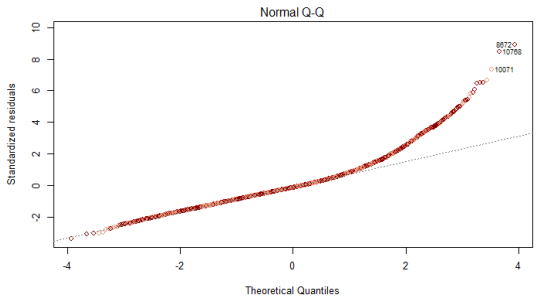
Objetiva-se seguidamente efetuar uma análise dos resíduos dos dois modelos de regressão linear construídos, por forma a verificar os pressupostos assumidos para os erros inerentes ao modelo, e, por conseguinte, inferir quanto à adequação do mesmo. A análise dos resíduos compreende a análise da normalidade, da autocorrelação e da homogeneidade das variâncias dos mesmos.

Os modelos 1 e 2 partilham as mesmas conclusões no que respeita à análise dos três pressupostos inerentes aos seus resíduos, porém serão só apresentados os resultados referentes ao modelo 2 por este ter sido o escolhido no presente estudo.

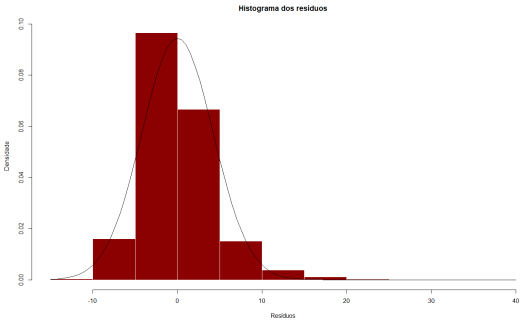
(a) **Normalidade dos resíduos**

Quanto à normalidade dos resíduos tanto os QQ-*plots* e histogramas obtidos, como os testes de normalidade implementados, coadunam-se a inexistência de normalidade nos dados. Os QQ-*plots* passíveis de serem observados na figura 5.12a, permitem constatar que os resíduos se encontram afastados da linha dos quantis teóricos. Da mesma forma, assimetria ao nível do histograma dos resíduos confirmam os resultados obtidos pelo QQ-*plot*. Relativamente aos testes de normalidade implementados é plausível constatar que os $p - values$ obtidos por aplicação dos testes de hipóteses de Anderson-Darling e Kolmogorov-Smirnov com correção de Lilliefors, são inferiores a 0.05, o que, a um nível de significância de 5% permite rejeitar a hipótese de normalidade dos resíduos. Os resultados obtidos encontram-se sintetizados na tabela 5.14. Note-se que o teste de Shapiro-Wilk não é aplicável visto que apenas só se deverá utilizar em conjuntos de

dados que contêm entre 3 e 5000 observações, sendo que o presente conjunto de dados possui um maior número de observações.




(a) QQ-*plot* dos resíduos do modelo 2




(b) Histograma dos resíduos do modelo 2

Figura 5.12: Gráficos referentes aos resíduos

Tabela 5.14: Resultados do modelo 2 referentes aos testes de normalidade

| | Anderson-Darling | KS com correção de Lilliefors | Decisão |
|---|---|---|---|
| Modelo 2 | $< 2.2e-16$ | $< 2.2e-16$ | Rejeita H0 |

Nota: $p\text{-}value \leq 0.05$ Rejeita H0; $p\text{-}value > 0.05$ Não Rejeita H0

A inexistência de normalidade dos resíduos do modelo de regressão linear implica a impossibilidade de calcular intervalos de confiança, inerentes a possíveis previsões associadas.

(b) **Autocorrelação**

Relativamente à autocorrelação dos resíduos pressupõe-se que os resíduos do modelo de regressão linear sejam independentes (não correlacionados). Assim, por forma a confirmar a presente suposição, foram implementados três testes, sendo eles o teste de Box-Pierce, o teste de Ljung-Box e o teste de Durbin Watson, cujos resultados se encontram sintetizados na tabela 5.15. A aplicação simultânea dos três testes permite comprovar as conclusões obtidas por cada um dos testes.

Tabela 5.15: Resultados do modelo 2 referentes aos testes de autocorrelação

| | Box-Pierce | Ljung-Box | Durbin Watson | Decisão |
|---|---|---|---|---|
| Modelo 2 | $< 2.2e-16$ | $< 2.2e-16$ | $< 2.2e-16$ | Rejeita H0 |

Nota: $p\text{-}value \leq 0.05$ Rejeita H0; $p\text{-}value > 0.05$ Não Rejeita H0

Os testes anteriores permitem decidir a favor da existência de correlação dos resíduos, sendo que essa decisão depende dos valores do $p - value$ reduzidos (inferiores 0.05), o que, a um nível de significância de 5%, permite rejeitar a hipótese nula, que assume a ausência de autocorrelação dos resíduos.

(c) **Heteroscedasticidade dos resíduos**

Um dos pressupostos inerentes aos resíduos do modelo de regressão é a heteroscedasticidade, a qual pode ser analisada empiricamente recorrendo a gráficos ou testada por testes de hipóteses. Relativamente ao gráfico que se encontra na figura 5.13, é possível visualizar "V" na horizontal, o que permite inferir quanto à existência de heteroscedasticidade das variâncias dos resíduos.

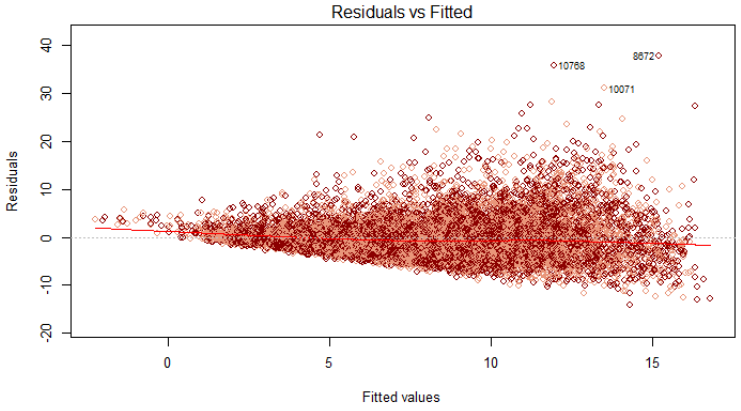



Figura 5.13: Gráfico para avaliar a heteroscedasticidade dos resíduos do modelo 2

Uma análise gráfica da figura 5.14 à representação dos resíduos em função de cada uma das variáveis explicativas, permitiu verificar também a presença de um "V"na horizontal (em particular no que diz respeito às variáveis cota, N_fr e dias_pp), o que permite supor a existência de heteroscedasticidade nos resíduos.

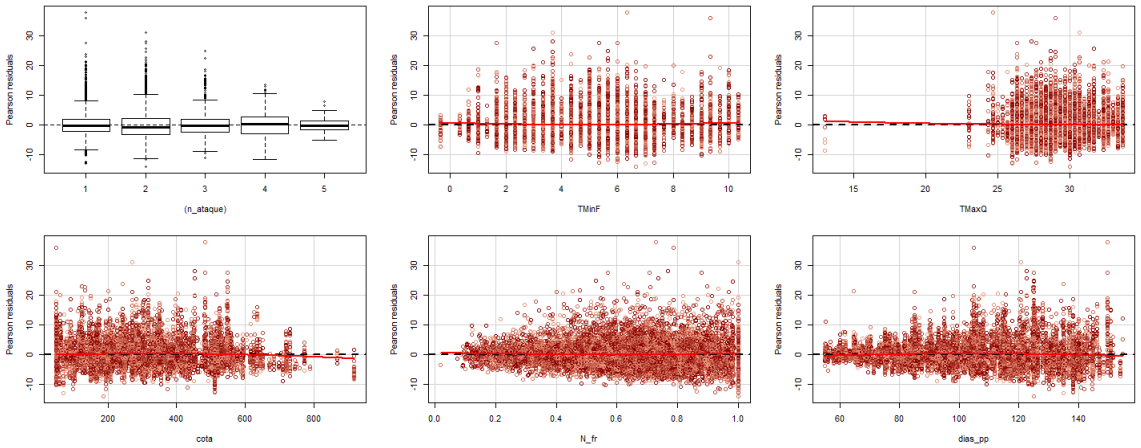



Figura 5.14: Gráficos para avaliar a heteroscedasticidade dos resíduos de Pearson (estandardizados) em função das variáveis, cota, N_fr e dias_pp, respetivamente

Com o intuito de sustentar a conjetura sugerida pela análise gráfica, implementou-se o teste de validação global das suposições de modelos lineares (*Global validation of linear model assumptions*), cuja função no *software* R é *gvlma*() (observável na tabela 5.16). A referida conjetura pode ser ainda comprovada com base em testes de hipóteses de deteção da heteroscedasticidade, nomeadamente o teste de Harrison-McCabe e o teste de Goldfeld-Quandt, cujos resultados se encontram sintetizados na tabela 5.17. Note-se que os testes de White e Breuch-Pagan não podem ser implementados pelo facto de os resíduos não apresentarem normalidade, tendo em conta que os mesmos têm apenas validade assintótica, pelo que as conclusões poderiam não ser fidedignas. A conclusão inerente ao teste da heteroscedasticidade coaduna-se com a conclusão gráfica, visto que o $p-value$ obtido foi inferior a 0.05 (observável nas tabelas 5.16 e 5.17).

Tabela 5.16: Resultados da função *gvlma*()

| | Value | p-value | Decision |
|---|---|---|---|
| Global Stat | 12754.81 | $0.000e+00$ | Assumptions NOT satisfied! |
| Skewness | 3264.24 | $0.000e+00$ | Assumptions NOT satisfied! |
| Kurtosis | 9405.77 | $0.000e+00$ | Assumptions NOT satisfied! |
| Link Function | 23.05 | $1.578e-06$ | Assumptions NOT satisfied! |
| Heteroscedasticity | 61.75 | $3.886e-15$ | Assumptions NOT satisfied! |

Tabela 5.17: Resultados do modelo 2 referentes aos testes de heteroscedasticidade

| | Harrison-McCabe | Goldfeld-Quandt | Decisão |
|---|---|---|---|
| Modelo 2 | $< 2.2e-16$ | $1.339e-07$ | Rejeita H0 |

Nota: *p-value*$\leq$0.05 Rejeita H0; *p-value*$>$0.05 Não Rejeita H0

Tendo em conta o explanado nesta secção verifica-se que apesar da inexistência de colinearidade nas variáveis independentes os restantes pressupostos ao nível dos resíduos do modelo não são validados. Por conseguinte, é plausível inferir que o modelo de regressão linear construído não se adequa ao conjunto de dados em estudo.

Assim, tendo como base o modelo 2 e no sentido de colmatar este problema, foram desenvolvidos três modelos de regressão linear múltipla que consistem na implementação de outras metodologias tais como a aplicação de transformação da Box-Cox aos dados e estimação na presença de autocorrelação (secção 5.4.2), a técnica HAC cuja aplicação resulta na alteração da matriz de variâncias e covariâncias (secção 5.4.3) e a regressão robusta (secção 5.4.4).

### 5.4.2 Modelo RLM 1

Nesta secção começa-se com a aplicação da transformação de Box-Cox aos dados originais com o intuito de corrigir as falhas do modelo construído anteriormente. Após a transformação dos dados é implementado um novo modelo de regressão linear e verificados os seus pressupostos. Uma vez que a transformação de Box-Cox não permitiu corrigir os dados de forma a tornar válidos todos os pressupostos foi ainda realizada uma transformação ao modelo de forma a ser possível a estimação na presença de autocorrelação.

#### Transformação de Box-Cox

Como já se referiu, aplicou-se a transformação de Box-Cox ao conjunto de dados em estudo, sendo o principal objetivo desta transformação garantir o pressuposto de normalidade e esse pressuposto não ser validado no modelo de regressão linear implementado, (Li, 2005). Porém, como já foi mencionado aquando do enquadramento teórico, esta transformação também pode ser útil para colmatar as restantes falhas associados aos resíduos dos modelos. A aplicação da presente metodologia fez uso apenas do modelo 2, visto que o primeiro modelo não foi considerado adequado pelas razões explanadas anteriormente.

A aplicação da presente metodologia utilizou a função *BoxCoxTrans()* da *package caret()* do *software* R, a qual permitiu determinar qual a melhor transformação dos dados estimando o valor de lambda mais adequado. Os resultados obtidos pela sua aplicação encontram-se no output 5.1.

```
> distBCMod3 <- BoxCoxTrans(DadosTotal$AMA_U_Proj12)
Box-Cox Transformation

12096 data points used to estimate Lambda

Input data summary:
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.1097  4.5590  7.4681  8.4778 11.2316 52.9177

Largest/Smallest: 482
Sample Skewness: 1.32

Estimated Lambda: 0.3
```

*Output* 5.1: Output da aplicação da transformação de Box-Cox

Salienta-se que a implementação da transformação de Box-Cox foi aplicada ao nível da variável dependente AMA_U_Proj12. Após a aplicação desta metodologia aos dados foi criada uma nova variável, AMA_New, sobre a qual foi aplicada regressão linear, fazendo uso da metodologia OLS. Os resultados obtidos por aplicação de regressão aos dados transformados encontram-se sintetizados na tabela 5.18.

Tabela 5.18: Resultados da estimação OLS dos dados transformados pela transformação de Box-Cox

| | *Variável dependente:* |
|---|---|
| | AMA_new |
| n_ataque 2 | −0.154*** |
| | (0.025) |
| n_ataque 3 | −0.464*** |
| | (0.028) |
| n_ataque 4 | −0.689*** |
| | (0.040) |
| n_ataque 5 | −2.403*** |
| | (0.074) |
| TMinF | 0.047*** |
| | (0.005) |
| TMaxQ | −0.024*** |
| | (0.004) |
| N_fr | 2.532*** |
| | (0.042) |
| cota | −0.00003 |
| | (0.0001) |
| dias_pp | 0.017*** |
| | (0.0005) |
| Constant | 0.155 |
| | (0.151) |
| Observations | 12,096 |
| $R^2$ | 0.432 |
| Adjusted $R^2$ | 0.432 |
| Residual Std. Error | 0.893 (df = 12086) |
| F Statistic | 1,021.923*** (df = 9; 12086) |
| Nota: | *p<0.1; **p<0.05; ***p<0.01 |

A aplicação da transformação de Box-Cox aos dados resultou na obtenção de $R^2$ superior ao obtido anteriormente, tendo sido na ordem de 0.4317. Relativamente aos valores de AIC e BIC os resultados alcançados encontram-se na ordem de 31611.26 e 31692.67, respetivamente. Quanto à estatística F o valor resultante é superior aos valores obtidos anteriormente, sendo o valor na ordem dos 1021.923. Tal conclui a maior adequação deste modelo aos dados.

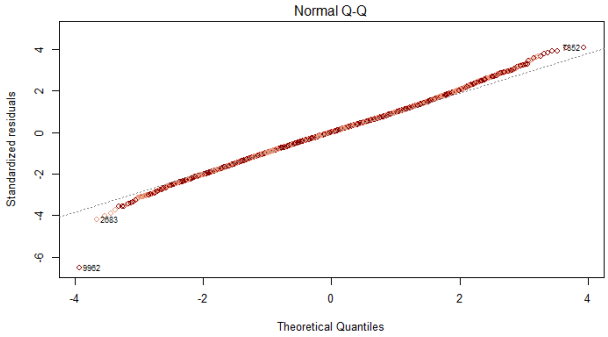
Tendo em consideração os testes de significância dos estimadores de regressão obtidos no modelo de regressão aplicados aos dados transformados é possível concluir que, de entre as variáveis explicativas utilizadas, apenas a variável cota não é significativamente diferente de zero. Com efeito, foi aplicada a metodologia *Stepwise* a qual permitiu remover esta variável. Após a remoção da variável foi realizada nova estimação OLS e os resultados apresentam-se na tabela 5.19.

Tabela 5.19: Resultados da estimação OLS dos dados transformados

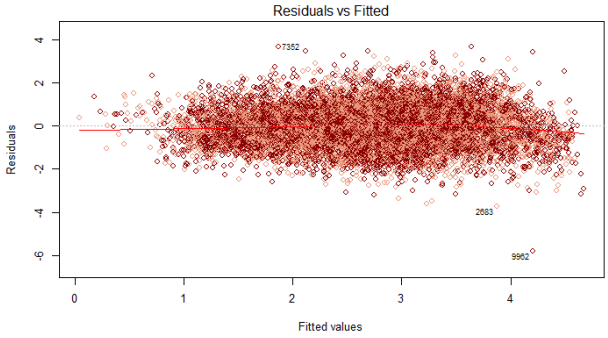
| | *Variável dependente:* |
|---|---|
| | AMA_new |
| n_ataque 2 | $-0.158^{***}$ |
| | (0.023) |
| n_ataque 3 | $-0.471^{***}$ |
| | (0.025) |
| n_ataque 4 | $-0.696^{***}$ |
| | (0.038) |
| n_ataque 5 | $-2.412^{***}$ |
| | (0.072) |
| TMinF | $0.048^{***}$ |
| | (0.005) |
| TMaxQ | $-0.024^{***}$ |
| | (0.004) |
| N_fr | $2.532^{***}$ |
| | (0.042) |
| dias_pp | $0.017^{***}$ |
| | (0.0005) |
| Constant | 0.146 |
| | (0.150) |
| Observations | 12,096 |
| $R^2$ | 0.432 |
| Adjusted $R^2$ | 0.432 |
| Residual Std. Error | 0.893 (df = 12087) |
| F Statistic | $1{,}149.697^{***}$ (df = 8; 12087) |
| Nota: | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

A remoção da variável cota levou a que o valor de $R^2$ se mantivesse igual ao do modelo

anterior, e a obtenção de valores de $p - value$ significativos para todas as variáveis independentes. No âmbito dos valores referentes ao AIC e BIC, constatou-se que estes são bastantes semelhantes aos obtidos no modelo anterior, porém ligeiramente melhores, tendo sido conseguidos os valores 31609.54 e 31683.55, respetivamente. Relativamente ao valor da estatística F obtido verificou-se o seu aumento, o que leva a conclusões análogas às referidas anteriormente.




(a) QQ-*plot* dos resíduos do modelo 2 com a transformação Box-Cox




(b) Gráfico para avaliar a heteroscedasticidade dos resíduos do modelo 2 com transformação Box-Cox

Figura 5.15: Gráficos de análise dos resíduos com transformação com Box-Cox

A análise dos resíduos baseia-se na análise da normalidade, da autocorrelação e da heteroscedasticidade dos resíduos. Primeiramente, efetua-se a análise da normalidade dos resíduos com base na observação gráfica do QQ-*plot* (figura 5.15a) e, por forma a comprovar os resultados obtidos, implementam-se testes de normalidade. A observação gráfica permite intuir a existência de elevada proximidade entre os pontos representados a vermelho e a reta de declive unitário, os quais representam, respetivamente, os valores empíricos da amostra, e os valores teóricos da distribuição normal. Assim, a observação gráfica permite conjeturar a existência de normalidade nos resíduos do modelo de regressão. No entanto, a aplicação dos testes contradiz a conjetura formulada, na medida em que a obtenção de valores do *p-value* reduzidos, implica a rejeição da normalidade nos resíduos. Os resultados encontram-se sintetizados na

tabela 5.20. Salienta-se que tal pode resultar da elevada dimensionalidade associada ao conjunto de dados em estudo, pelo que os testes de normalidade podem falhar, (Hall et al., 2011). Porém se se considerar válido o resultado dos testes de hipóteses apresentados na tabela 5.20, conclui-se a inexistência de normalidade, pelo que este pressuposto é violado. Sendo conhecida a elevada dimensão do conjunto de dados e a análise do QQ-*plot*, vai-se admitir a existência de normalidade.

Tabela 5.20: Resultados referentes aos testes de normalidade após a transformação de Box-Cox

| | Anderson-Darling | KS com correção de Lilliefors | Decisão |
|---|---|---|---|
| p-value | $3.152e-09$ | $6.523e-05$ | Rejeita H0 |

Nota: *p-value*$\leq$0.05 Rejeita H0; *p-value*$>$0.05 Não Rejeita H0

Relativamente à autocorrelação dos resíduos, a aplicação dos testes de autocorrelação levam à rejeição da hipótese nula, o que implica concluir a existência da mesma nos resíduos em estudo (ver tabela 5.21).

Tabela 5.21: Resultados referentes aos testes de autocorrelação após a transformação Box-Cox

| | Box-Pierce | Ljung-Box | Durbin Watson | Decisão |
|---|---|---|---|---|
| p-value | $< 2.2e-16$ | $< 2.2e-16$ | $< 2.2e-16$ | Rejeita H0 |

Nota: *p-value*$\leq$0.05 Rejeita H0; *p-value*$>$0.05 Não Rejeita H0

Os resultados obtidos por aplicação do *gvlma*() permitem supor a existência de homogeneidade das variâncias dos resíduos neste modelo bem como a existência de simetria na curva da função densidade de probabilidade da distribuição empírica (tabela 5.22). Os testes de heteroscedasticidade e a representação gráfica corroboram o explanado anteriormente (tabela 5.23 e figura 5.15b). De forma análoga ao explicado anteriormente a implementação dos testes de White e Breuch-Pagan podem induzir a conclusões erradas, visto que a sua aplicação não é adequada em resíduos que não satisfaçam o princípio de normalidade.

Tabela 5.22: Resultados da função gvlma() após a transformação Box-Cox

| | Value | p-value | Decision |
|---|---|---|---|
| Global Stat | 192.61716 | 0.0000 | Assumptions NOT satisfied! |
| Skewness | 0.05912 | 0.8079 | Assumptions acceptable. |
| Kurtosis | 115.67228 | 0.0000 | Assumptions NOT satisfied! |
| Link Function | 76.83662 | $1.578e-06$ | Assumptions NOT satisfied! |
| Heteroscedasticity | 0.04914 | 0.8246 | Assumptions acceptable. |

Tabela 5.23: Resultados referentes aos testes de heteroscedasticidade após a transformação de Box-Cox

| | Harrison-McCabe | Goldfeld-Quandt | Decisão |
|---|---|---|---|
| p-value | 0.868 | 0.8777 | Não Rejeita H0 |

Nota: *p-value*≤0.05 Rejeita H0; *p-value*>0.05 Não Rejeita H0

Com esta transformação, considerando os gráficos referentes à normalidade e o que se sabe quanto a conjuntos de dados com elevada dimensão, é possível considerar que esta transformação foi profícua na validação do pressuposto da normalidade, isto é, estamos agora perante a existência de normalidade. Foi também possível eliminar a heteroscedasticidade existente no modelo sem transformação. No entanto, não se conseguiu contornar a problemática da existência de autocorrelação dos resíduos. Então prossegue-se com a modelação dos resíduos e a sua inclusão no modelo.

### Métodos de estimação na presença de autocorrelação

No processo de modelação dos resíduos na presença de autocorrelação fez-se uso do método de Cochrane-Orcutt–CO e do método de Durbin, sendo que o primeiro foi implementado no *software* R com auxilio da função *cochrane.orcutt()* do *package orcutt*, já para o segundo recorreu-se ao *software* E.Views.

A implementação do método de Cochrane-Orcutt produziu os resultados apresentados na tabela 5.24.

Tabela 5.24: Resultados do método de Cochrane-Orcutt

| | *Variável dependente:* |
|---|---|
| | AMA_new |
| n_ataque 2 | −0.144395*** |
| | (0.031) |
| n_ataque 3 | −0.534961*** |
| | (0.035) |
| n_ataque 4 | −0.712412*** |
| | (0.051) |
| n_ataque 5 | −2.358530*** |
| | (0.107) |
| TMinF | 0.027154*** |
| | (0.006) |
| TMaxQ | −0.006963* |
| | (0.004) |
| N_fr | 2.528063*** |
| | (0.043) |
| dias_pp | 0.017467*** |
| | (0.001) |
| Constant | −0.329910** |
| | (0.142) |
| $\rho$ | 0.508428 |
| Interactions | 5 |
| Observations | 12,096 |
| $R^2$ | 0.3129 |
| Adjusted $R^2$ | 0.3126 |
| Residual Std. Error | 0.7709 (df = 12089) |
| F Statistic | 687.9*** (df = 5; 12089) |
| Nota: | *p<0.1; **p<0.05; ***p<0.01 |

Note-se que a aplicação do método de CO, permitiu verificar que o valor da autocorrelação existente nos resíduos tem um valor que deve ser tido em linha de conta, pelo que é completamente adequado ele ser incluído no modelo de regressão ajustado.

O modelo transformado que resulta da aplicação deste método é o seguinte:

$$\begin{aligned} AMA\_new_i = {} & \beta_0(1-\rho) + \rho AMA\_new_{i-1} + \beta_1(n\_ataque2_i - \rho n\_ataque2_{i-1}) \\ & + \beta_2(n\_ataque3_i - \rho n\_ataque3_{i-1}) + \beta_3(n\_ataque4_i - \rho n\_ataque4_{i-1}) \\ & + \beta_4(n\_ataque5_i - \rho n\_ataque5_{i-1}) + \beta_5(TMinF_i - \rho TMinF_{i-1}) \\ & + \beta_6 TMaxQ_i - \rho TMaxQ_{i-1}) + \beta_7(N\_fr_i - \rho N\_fr_{i-1}) \\ & + \beta_8(dias\_pp_i - \rho dias\_pp_{i-1}) + u_i. \end{aligned}$$

Por substituição dos coeficientes de regressão estimados e do $\rho$ presentes na tabela 5.24 na equação do modelo anterior, tem-se o seguinte modelo

$$\begin{aligned} AMA\_new_i = {} & -0.330(1-0.508) + 0.508 AMA\_new_{i-1} - 0.144(n\_ataque2_i \\ & - 0.508 n\_ataque2_{i-1}) - 0.535(n\_ataque3_i - 0.508 n\_ataque3_{i-1}) \\ & - 0.712(n\_ataque4_i - 0.508 n\_ataque4_{i-1}) - 2.359(n\_ataque5_i \\ & - 0.508 n\_ataque5_{i-1}) + 0.027(TMinF_i - 0.508 TMinF_{i-1}) \\ & - 0.007(TMaxQ_i - 0.508 TMaxQ_{i-1}) + 2.528(N\_fr_i - 0.508 N\_fr_{i-1}) \\ & + 0.017(dias\_pp_i - 0.508 dias\_pp_{i-1}). \end{aligned}$$

Salienta-se que este método também foi implementado com recurso ao *software E.Views*, em que os resultados coincidiram com os obtidos pelo *software* R. A diferença entre os dois *softwares* é que o R só precisa de um passo para a obtenção do modelo, pois ele intrinsecamente realiza os dois passos inerentes a este método, já o *software E.Views* necessita que se efetuem dois passos manualmente.

Seguidamente foi aplicado método de Durbin no *software E.Views*, em que o primeiro passo, ou seja, a obtenção de $\rho$, se encontra no anexo C.2, já os resultados referentes ao passo 2 se encontram sintetizados na tabela 5.25.

Tabela 5.25: Resultados do passo 2 do método de Durbin

Dependent Variable: AMA_NEW-0.506413*AMA_NEW(-1)
Method: Least Squares
Sample (adjusted): 2 12096
Included observations: 12095 after adjustments

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | −0.162235 | 0.070121 | −2.313643 | 0.0207 |
| (N_ATAQUE=2)-(N_ATAQUE(-1)=2)*0.506413 | −0.144517 | 0.031402 | −4.602187 | 0.0000 |
| (N_ATAQUE=3)-(N_ATAQUE(-1)=3)*0.506413 | −0.534591 | 0.034809 | −15.35787 | 0.0000 |
| (N_ATAQUE=4)-(N_ATAQUE(-1)=4)*0.506413 | −0.712331 | 0.050987 | −13.97071 | 0.0000 |
| (N_ATAQUE=5)-(N_ATAQUE(-1)=5)*0.506413 | −2.359154 | 0.107034 | −22.04113 | 0.0000 |
| TMINF-0.506413*TMINF(-1) | 0.027302 | 0.005961 | 4.580130 | 0.0000 |
| TMAXQ-0.506413*TMAXQ(-1) | −0.007035 | 0.003659 | −1.922648 | 0.0545 |
| N_FR-0.506413*N_FR(-1) | 2.528129 | 0.042692 | 59.21756 | 0.0000 |
| DIAS_PP-0.506413*DIAS_PP(-1) | 0.017468 | 0.000701 | 24.91249 | 0.0000 |

| | | | |
|---|---|---|---|
| R-squared | 0.313426 | Mean dependent var | 1.351576 |
| Adjusted R-squared | 0.312971 | S.D. dependent var | 0.930159 |
| S.E. of regression | 0.770983 | Akaike info criterion | 2.318442 |
| Sum squared resid | 7184.088 | Schwarz criterion | 2.323949 |
| Log likelihood | −14011.78 | Hannan-Quinn criter. | 2.320288 |
| F-statistic | 689.6677 | Durbin-Watson stat | 2.250580 |
| Prob(F-statistic) | 0.000000 | | |

O modelo que resulta após a aplicação do método está descrito na equação seguinte:

$$\begin{aligned} AMA\_new_i = {} & -1.622(1-0.506) + 0.506 AMA\_new_{i-1} - 0.144(n\_ataque2_i \\ & - 0.506 n\_ataque2_{i-1}) - 0.535(n\_ataque3_i - 0.506 n\_ataque3_{i-1}) \\ & - 0.712(n\_ataque4_i - 0.506 n\_ataque4_{i-1}) \\ & - 2.359(n\_ataque5_i - 0.506 n\_ataque5_{i-1}) \\ & + 0.027(TMinF_i - 0.506 TMinF_{i-1}) - 0.007(TMaxQ_i - 0.506 TMaxQ_{i-1}) \\ & + 2.528(N\_fr_i - 0.506 N\_fr_{i-1}) + 0.017(dias\_pp_i - 0.506 dias\_pp_{i-1}). \end{aligned}$$

Note-se que outra maneira de obter os coeficientes de regressão na presença de autocorrelação é acrescentando no modelo de regressão no *software E.Views* a instrução AR(p), isto é, o modelo autorregressivo de ordem p.

Nas secções seguintes ir-se-ão aplicar outras técnicas que também consigam contornar estes problemas inerentes aos resíduos, para assim conseguirmos ter um modelo adequado ao problema em estudo.

### 5.4.3 Modelo RLM 2

Nesta secção ir-se-á aplicar outra metodologia de análise por forma a colmatar as falhas inerentes ao modelo 2, baseada nas matrizes de variâncias e covariâncias consistentes na presença de heteroscedasticidade e/ou autocorrelação.

Esta técnica é aplicada a nível da matriz de variâncias e covariâncias, por forma a tornar os estimadores consistentes na presença de heteroscedasticidade e autocorrelação. Os resultados obtidos são apresentados na tabela 5.26, os quais foram conseguidos a partir da função *vcovHAC()* do *package sandwich*.

Tabela 5.26: Resultados da regressão com a aplicação de HAC

| | *Variável dependente:* | |
|---|---|---|
| | AMA_U_Proj12 | |
| | OLS (1) | HAC (2) |
| n_ataque 2 | −0.876*** | −0.876** |
| | (0.117) | (0.373) |
| n_ataque 3 | −2.526*** | −2.526*** |
| | (0.132) | (0.387) |
| n_ataque 4 | −3.562*** | −3.562*** |
| | (0.189) | (0.582) |
| n_ataque 5 | −9.345*** | −9.345*** |
| | (0.351) | (0.684) |
| TMinF | 0.143*** | 0.143** |
| | (0.024) | (0.064) |
| TMaxQ | −0.129*** | −0.129*** |
| | (0.019) | (0.033) |
| cota | 0.001** | 0.001 |
| | (0.0003) | (0.001) |
| N_fr | 9.966*** | 9.966*** |
| | (0.197) | (0.475) |
| dias_pp | 0.076*** | 0.076*** |
| | (0.002) | (0.007) |
| Constant | −1.501** | −1.501 |
| | (0.716) | (1.201) |
| Observations | 12,096 | 12,096 |
| $R^2$ | 0.366 | 0.366 |
| Adjusted $R^2$ | 0.365 | 0.365 |
| Residual Std. Error (df = 12086) | 4.227 | 4.227 |
| F Statistic (df = 9; 12086) | 773.733*** | 773.733*** |
| Nota: | *p<0.1; **p<0.05; ***p<0.01 | |

A implementação da presente metodologia implicou apenas alterações ao nível dos erros padrão e $p-value$, sendo que os restantes valores inerentes ao modelo de regressão se mantém iguais. Tal coaduna-se com as informações teóricas apresentadas na secção 4.5.

Em relação ao valor do $R^2$ obtido por aplicação de HAC sob o modelo de regressão linear, este valor é igual ao obtido por aplicação do OLS, visto que os coeficientes de regressão não se alteram entre ambos.

As conclusões relativas à análise global de significância do modelo de regressão são análogas às referidas anteriormente para o modelo de regressão OLS, uma vez que o valor da estatística F se mantém igual.

O modelo que resulta após a aplicação dos métodos traduz-se na seguinte equação:

$$AMA\_U\_Proj12 = -1.501 - 0.876Ataque_2 - 2.526Ataque_3 - 3.562Ataque_4 - 9.345Ataque_5 + 0.143TMinF - 0.129TMaxQ - 0.001cota + 9.966N\_fr + 0.076dias\_pp \quad (5.4)$$

Apesar desta técnica ser aplicável na situação em que se verifica a falha dos pressupostos de regressão linear, com o intuito de analisar metodologias alternativas, que permitam de igual forma contornar a violação dos pressupostos de regressão linear, seguidamente foi desenvolvido um modelo de regressão robusta.

### 5.4.4 Modelo RLM 3

Outro modelo de regressão linear múltiplo alternativo aos dois anteriores o que resulta da aplicação de regressão linear múltipla robusta – RML aos dados. Este não só se justifica tendo por base a violação dos pressupostos do OLS mas também pela existência de *outliers* no conjunto de dados. Salienta-se que a presente metodologia, tal como a abordagem apresentada na secção 5.4.3, não resolve as falhas dos pressupostos nos resíduos do modelo, tornando apenas os estimadores consistentes.

Desta forma, segue-se uma análise aos *outliers* presentes no conjunto de dados em estudo.

#### Deteção de *outliers* e observações influentes

A análise referente à deteção dos *outliers* e observações influentes é feita sobre o modelo 2 pelo explanado anteriormente.

Por forma a detetar as observações atípicas, isto é, os *outliers*, procedeu-se ao cálculo dos resíduos studentizados. Em primeira análise recorreu-se ao *boxplot* dos resíduos representado na figura 5.16, de onde se concluiu a existência de *outliers*.

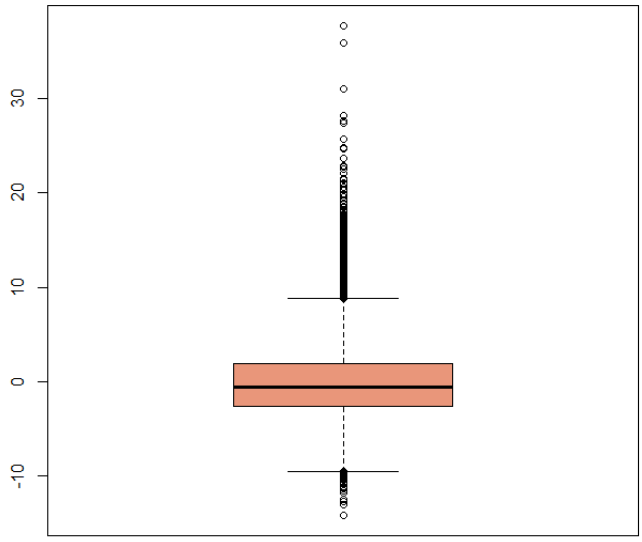



Figura 5.16: *Boxplot* dos resíduos associados ao modelo de regressão linear

Uma avaliação aos resíduos de Pearson estandardizados permitiu concluir que aproximadamente 4.94% observações estão fora do intervalo de confiança entre -1.96 a 1.96, e, por conseguinte são considerados *outliers*. O que coincide com os resultados obtidos na análise ao *box-plot* anterior.

Por fim, realizou-se uma análise à distância de Cook (considerando-se um ponto de corte de $DC > \frac{4}{n-k-1} = 0.00033$). Com base na distancia de Cook verificou-se que cerca de 4.86% são observações influentes; o que pode ser visualizado através do gráfico das distâncias de Cook e o respetivo ponto de corte, apresentado na figura 5.17.

Os gráficos das várias medidas de diagnóstico utilizadas e os gráficos referentes ao *leverage* para cada variável independente em função da variável explicativa encontram-se no anexo C.3.

Uma análise gráfica à figura 5.17 permite destacar seis *outliers*, sendo estes os mais discrepantes relativamente aos restantes.

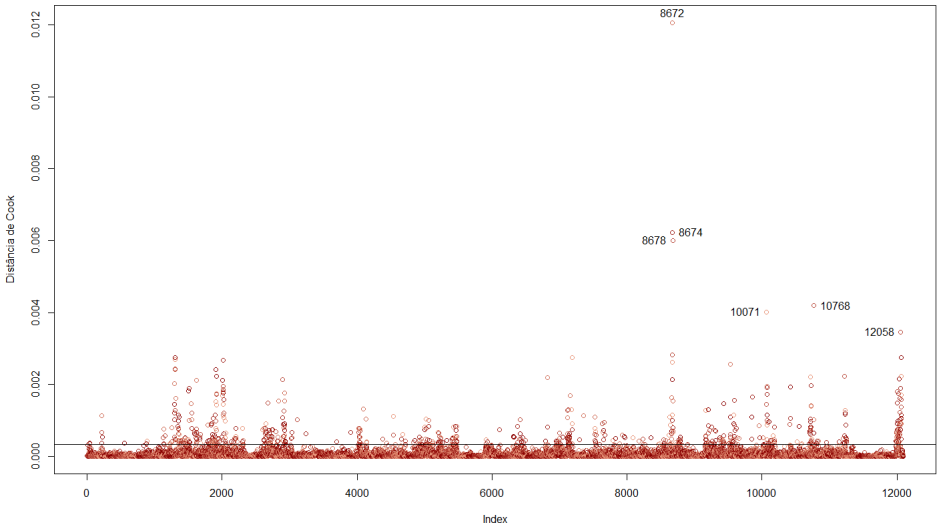



Figura 5.17: Distância de Cook e o ponto de corte considerado

### Regressão linear múltipla robusta – RLMR

A existência de *outliers* e de observações influentes aliada à violação dos pressupostos do OLS justifica a aplicação de uma abordagem de regressão robusta.

Esta técnica permite aplicar diversos métodos para calcular os estimadores, tendo sido escolhidos os estimadores M e MM na presente investigação. Refere-se, por fim, que, tendo por base a literatura consultada, o método mais apropriado na presença de *outliers* é o uso de estimadores MM. A presente metodologia de regressão robusta faz uso do comando *rlm* do *software* R.

Seguidamente apresentam-se os resultados da aplicação do modelo de regressão robusta, utilizando os estimadores M (consultar tabela 5.27).

Tabela 5.27: Resultados da regressão com a aplicação dos estimadores M

| | *Variável dependente:* | |
|---|---|---|
| | AMA_U_Proj12 | |
| | OLS | Robust (Estimator M) |
| n_ataque 2 | −0.876*** | −0.745*** |
| | (0.117) | (0.100) |
| n_ataque 3 | −2.526*** | −1.976*** |
| | (0.132) | (0.113) |
| n_ataque 4 | −3.562*** | −2.845*** |
| | (0.189) | (0.162) |
| n_ataque 5 | −9.345*** | −8.160*** |
| | (0.351) | (0.301) |
| TMinF | 0.143*** | 0.197*** |
| | (0.024) | (0.021) |
| TMaxQ | −0.129*** | −0.129*** |
| | (0.019) | (0.017) |
| cota | 0.001** | −0.0002 |
| | (0.0003) | (0.0002) |
| N_fr | 9.966*** | 9.599*** |
| | (0.197) | (0.169) |
| dias_pp | 0.076*** | 0.063*** |
| | (0.002) | (0.002) |
| Constant | −1.501** | −0.504 |
| | (0.716) | (0.613) |
| Observations | 12,096 | 12,096 |
| $R^2$ | 0.366 | 0.355 |
| Adjusted $R^2$ | 0.365 | 0.355 |
| Residual Std. Error (df = 12086) | 4.227 | 3.302 |
| F Statistic | 773.733*** (df = 9; 12086) | |
| Nota: | | *p<0.1; **p<0.05; ***p<0.01 |

O valor $R^2$ obtido foi diferente dos demais, sendo que esse valor foi de 0.355.

Relativamente aos valores de AIC e BIC, a aplicação da presente metodologia permitiu obter os valores 69397.33 e 69471.34, respetivamente.

A expressão do modelo obtido pela aplicação do estimador–M, traduz-se na equação seguinte:

$$AMA\_U\_Proj12 = -0.504 - 0.745Ataque_2 - 1.976Ataque_3 - 2.845Ataque_4 - 8.160Ataque_5 + 0.143TMinF - 0.129TMaxQ - 0.001cota + 9.966N\_fr + 0.076dias\_pp \quad (5.5)$$

Sendo o estimador–MM mais eficiente, nomeadamente, na presença de *outliers*, aplicou-se seguidamente a metodologia de regressão robusta, utilizando para tal os estimadores–MM. Os resultados obtidos encontram-se sintetizados na tabela 5.28.

Tabela 5.28: Resultados da regressão com a aplicação dos estimadores MM

| | *Variável dependente:* | |
|---|---|---|
| | AMA_U_Proj12 | |
| | *OLS* | *Robust (Estimator MM)* |
| n_ataque 2 | −0.876*** | −0.634*** |
| | (0.117) | (0.098) |
| n_ataque 3 | −2.526*** | −1.656*** |
| | (0.132) | (0.111) |
| n_ataque 4 | −3.562*** | −2.449*** |
| | (0.189) | (0.159) |
| n_ataque 5 | −9.345*** | −7.548*** |
| | (0.351) | (0.294) |
| TMinF | 0.143*** | 0.226*** |
| | (0.024) | (0.020) |
| TMaxQ | −0.129*** | −0.128*** |
| | (0.019) | (0.016) |
| cota | 0.001** | −0.001*** |
| | (0.0003) | (0.0002) |
| N_fr | 9.966*** | 9.364*** |
| | (0.197) | (0.165) |
| dias_pp | 0.076*** | 0.055*** |
| | (0.002) | (0.002) |
| Constant | −1.501** | 0.082 |
| | (0.716) | (0.600) |
| Observations | 12,096 | 12,096 |
| $R^2$ | 0.366 | 0.343 |
| Adjusted $R^2$ | 0.365 | 0.342 |
| Residual Std. Error (df = 12086) | 4.227 | 3.211 |
| F Statistic | 773.733*** (df = 9; 12086) | |
| Nota: | | *p<0.1; **p<0.05; ***p<0.01 |

A aplicação de regressão robusta com os estimadores MM permitiu obter valores de AIC e BIC, respetivamente, de 69604.07 e 69678.08. Relativamente ao $R^2$, verifica-se que a função implementada no *software* R para a obtenção deste modelo não apresenta o seu valor. Desta forma procedeu-se ao cálculo deste valor recorrendo, para tal, à função também aplicada ao modelo do estimador M. O valor obtido foi 0.343.

O modelo obtido após a anterior metodologia, expressa-se na seguinte equação:

$$AMA\_U\_Proj12 = 0.082 - 0.634Ataque_2 - 1.656Ataque_3 - 2.449Ataque_4 - 7.548Ataque_5 + 0.226TMinF - 0.128TMaxQ - 0.001cota + 9.364N\_fr + 0.055dias\_pp \quad (5.6)$$

### 5.4.5 Análise comparativa dos modelos de regressão construídos

No presente subcapítulo o objetivo é, através da comparação dos modelos ajustados nas secções anteriores, no que concerne aos valores de $R^2$, AIC e BIC obtidos, selecionar o modelo que melhor responde ao problema formulado na secção 5.1.

Tabela 5.29: Sintetização dos resultados relevantes dos dois modelos preliminares formulados

| | $R^2$ | AIC | BIC |
|---|---|---|---|
| OLS – Modelo 1 | 0.3656 | 69201.67 | 69297.88 |
| OLS – Modelo 2 | 0.3651 | 69210.09 | 69291.49 |

Na tabela 5.29 são apresentados os valores de comparação supracitados, entre os modelos 1 e 2 explanados anteriormente. Verifica-se que em ambos os modelos os resíduos não satisfazem os pressupostos de regressão linear, pelo que este fator não foi usado como desempate entre os modelos.

Porém, tendo por base o explanado nas secções anteriores, foi utilizado o modelo 2 na aplicação das diferentes técnicas de regressão aplicadas, cujos valores de $R^2$, AIC e BIC se encontram na tabela 5.30.

Tabela 5.30: Sintetização dos resultados relevantes do modelo 2 com as diferentes técnicas

| | | $R^2$ | AIC | BIC |
|---|---|---|---|---|
| | OLS | 0.3651 | 69210.09 | 69291.49 |
| | Box-Cox | 0.4317 | 31611.26 | 31692.67 |
| RLM 1 | CO | 0.3126 | — | — |
| RLM 1 | Durbin | 0.3130 | — | — |
| RLM 2 | HAC | 0.3651 | 69210.09 | 69291.49 |
| RLM 3 | rlm_M | 0.3547 | 69397.33 | 69471.34 |
| RLM 3 | rlm_MM | 0.3425 | 69604.07 | 69678.08 |

No que se refere à transformação de Box-Cox, explicada na secção 5.4.2, verificou-se que esta foi profícua para colmatar a heteroscedasticidade das variâncias dos resíduos do modelo. Relativamente à normalidade dos resíduos associados ao presente modelo transformado, verifica-se que graficamente, o QQ-*plot* sugere a existência de normalidade, visível pela proximidade entre os pontos dos resíduos e a reta relativa aos valores teóricos da distribuição Normal. No entanto, a aplicação de testes de normalidade aos resíduos deste modelo, permite concluir que os resíduos não seguem a distribuição Normal. Porém, esta conclusão é duvidosa na medida em que, literatura consultada, e tal como referido anteriormente, os testes de normalidade podem falhar na presença de inúmeras observações do conjunto de dados. Na presente investigação considerou-se, por estas razões, a existência de normalidade, (Hall et al., 2011). Relativamente à autocorrelação dos resíduos do presente modelo transformado, verifica-se que este pressuposto é violado, o que motivou a aplicação do método de Cochrane Orcutt e o método de Durbin. Assim, apesar da transformação de Box-Cox permitir a obtenção de valores de $R^2$ mais favoráveis, a violação dos pressupostos de regressão linear, em particular, no que diz respeito à normalidade e à autocorrelação dos resíduos, motivou a aplicação de métodos mais robustos de regressão, no entanto os métodos referidos originaram um $R^2$ mais baixo do que os restantes aplicados.

Por forma a colmatar a problemática de heteroscedasticidade e autocorrelação aplicou-se a metodologia HAC, em que se constata, tal como seria expectável, que os valores de $R^2$, AIC e BIC são análogos aos valores obtidos por aplicação de regressão linear pelo método OLS (modelo 2), visto que a referida metodologia, sendo aplicada sob o modelo 2, não altera os estimadores, apenas os torna consistentes face à violação dos pressupostos de regressão linear, nomeadamente, na presença de normalidade, autocorrelação e heteroscedasticidade dos resíduos.

De maneira a construir modelos de igual forma consistentes quando se verifica a falha dos pressupostos de regressão linear, foram aplicadas ainda metodologias de regressão robusta, '*rlm*', particularmente fazendo uso dos estimadores M e MM. A sua aplicação levou, em ambos os casos à obtenção, de um valor de $R^2$ mais reduzido, comparativamente com o método de regressão linear com OLS, enquanto que os valores de AIC e BIC são superiores aos do modelo OLS. Assim, apesar dos coeficientes de comparação de modelos serem considerados menos adequados, a construção de estimadores consistentes pela aplicação dos métodos supracitados, apresenta-se como vantajosa em detrimento do método OLS. Uma análise comparativa entre o uso de estimadores M e MM na regressão robusta permite observar que os valores das medidas de adequação dos modelos são bastante semelhantes entre si, apesar de, para os três valores, o estimador M apresentar vantagens, apesar de pouco significativas, em detrimento do estimador MM. De facto, os valores de $R^2$ são mais reduzidos no método de regressão MM, enquanto que os valores de AIC e BIC são mais reduzidos nos estimadores M. Porém, apesar de os estimadores M permitirem a obtenção de resultados mais favoráveis segundo as medidas de adequação consideradas, opta-se pela escolha dos estimadores MM, visto que estes, segundo literatura consultada, permitem a obtenção de melhores resultados na presença de *outliers* (característica associada ao conjunto de dado em estudo).

Note-se que também foram testados modelos de regressão linear na forma multiplicativa e mista no que diz respeito à inclusão da variável *dummy*, porém os resultados obtidos não foram considerados vantajosos para o estudo.

A titulo conclusivo, e no seguimento do explanado anteriormente, a escolha do modelo recai no modelo com estimadores MM, visto que, as diferenças das medidas de adequação dos modelos não são significativas, e atendendo a que este não é influenciado pela presença de *outliers*, permitindo construir estimadores consistentes na ausência de normalidade e presença de autocorrelação e heteroscedasticidade dos resíduos. Desta forma, seguidamente apresenta-se e analisa-se de forma detalhada o modelo construído pela metodologia supracitada.

$$AMA\_U\_Proj12 = 0.082 - 0.634Ataque_2 - 1.656Ataque_3 - 2.449Ataque_4 - 7.548Ataque_5 + 0.226TMinF - 0.128TMaxQ - 0.001cota + 9.364N\_fr + 0.055dias\_pp \quad (5.7)$$

Uma breve análise aos coeficientes de regressão das variáveis independentes do modelo de regressão apresentado (influência exercida sobre o AMA_U_Proj12) permite concluir que:

(i) Relativamente aos coeficientes de regressão da variável ataque existe um decréscimo cada vez maior no valor médio do AMA_U_Proj12 à medida que o nível de ataque aumenta, realçando que no nível de ataque com maior estragos (muito forte) verifica-se que sobre o valor médio da variável dependente apresenta um decréscimo acentuado de 7.548 unidades, isto quando todo o resto se mantém constante. Em relação à ausência de ataque, como seria previsto, o valor médio do AMA_U_Proj12 aumenta em 0.082 unidades.

(ii) Em relação à variável temperatura mínima dos três meses mais frios, TMinF, o coeficiente de regressão representa um aumento de 0.226 unidades no que se refere ao valor médio da variável AMA_U_Proj12, para um aumento unitário de TMinF, mantendo as restantes variáveis constantes.

(iii) Quanto à variável temperatura máxima dos três meses mais quentes, TMaxQ, contrariamente à anterior, o seu coeficiente de regressão denota uma diminuição de 0.128 unidades relativamente ao valor médio da variável dependente, AMA_U_Proj12, para um aumento unitário de TMaxQ, quando tudo o resto se mantém constante.

(iv) A respeito à variável independente cota constata-se que esta tem um efeito negativo na variável de resposta (AMA_U_Proj12), visto que, o coeficiente de regressão desta variável é de -0.001, o que traduz numa diminuição sobre o valor médio da variável AMA_U_Proj12 de 0.001, com todo o resto constante.

(v) No que concerne aos coeficientes de regressão das variáveis dias de precipitação (dias_pp) e número de árvores vivas (N_fr) representam uma influência positiva em relação ao AMA_U_Proj12, isto porque os seus coeficientes de regressão apresentam valores de 0.005 e 9.364, respetivamente. Relativamente a esta última variável o valor apresentado é elevado, isso já seria expectável pelo facto de que quantas mais árvores vivas mais o valor médio da variável AMA_U_Proj12 aumenta, isto é, um acréscimo de 1 unidade em N_fr espera-se que provoque um aumento médio de 9.364 unidades sobre o valor médio da variável AMA_U_Proj12, mantendo todo o resto constante. No que concerne à variável dias_pp, tal como seria expectável, esta variável permite incrementar positivamente a produtividade, visto que os eucaliptos são espécies que necessitam de elevadas quantidades de água.

Note-se que embora alguns dos valores referentes aos coeficientes de regressão se encontrem próximos de zero isso não significa que a variável deva ser removida do modelo. É necessário ter em consideração que as variáveis independentes poderão ter diferentes unidades de medida, e, por conseguinte, os respetivos coeficientes de regressão ordens de grandeza diferentes.

# Capítulo 6

# Conclusões

Primeiramente, deixo aqui o meu primordial agradecimento à entidade acolhedora, que me permitiu a realização do estágio curricular e da investigação que culminou na presente dissertação, elaborada no âmbito do Mestrado em Matemática e Aplicações, no ramo de Estatística e Otimização. Este trabalho permitiu-me tomar consciência da problemática inerente à atuação da espécie invasora, *Gonipterus platensis*. Assim, face às consequências nefastas na produtividade das parcelas associadas à sua atuação, é primordial desenvolver metodologias que permitam prever a produtividade das parcelas na presença da praga, já que a atuação do parasita é difícil de contornar. Com efeito, foi possível constatar a dificuldade associada ao combate da mesma, visto que as inúmeras técnicas desenvolvidas até ao momento, não se consideraram profícuas. Desta forma, prever a produtividade no sentido de tentar contornar as perdas inerentes à atuação da praga mostra-se como essencial na indústria papeleira. Nesta área salienta-se o uso maioritário da espécie *Eucalyptus globulus*, sendo a espécie onde a monitorização da praga incide na sua totalidade, visto que a monitorização do *Gonipterus platensis* foi realizada exclusivamente nesta espécie, não sendo consideradas outras espécies.

O trabalho desenvolvido permitiu construir um modelo capaz de prever a produtividade de uma parcela na presença de diversas variáveis que, constantemente influenciam a mesma. Assim, o modelo construído inclui fatores como a classificação do nível de ataque do *Gonipterus platensis*, as temperaturas mínimas dos três meses mais frios, as temperaturas máximas dos três meses mais quentes, a cota, a frequência relativa do número de árvores vivas, os dias de precipitação registados em cada parcela.

A variável dependente considerada no modelo foi o AMA_U_Proj12, visto que esta variável permite intuir quanto à produtividade das parcelas, sendo pautada de elevada importância para a entidade de acolhimento.

O modelo construído permitiu evidenciar que todos os níveis de ataque implicam o decrescimento da produtividade da parcela, sendo que o nível de ataque inexistente influencia positivamente a produtividade da parcela, tal como seria expectável. No que diz respeito a condições ambientais e climáticas, o modelo construído permitiu intuir que, caso as restantes

variáveis consideradas no modelo se mantenham constantes, o aumento da cota e o aumento das temperaturas máximas dos três meses mais quentes influenciam negativamente a produtividade da parcela, enquanto que o aumento das temperaturas mínimas dos três meses mais frios e o aumento do número de dias de precipitação, caso as restantes variáveis se mantenham constantes, implicam um crescimento da produtividade da parcela. Por fim, verifica-se que, por cada aumento unitário na frequência relativa de árvores vivas nas parcelas, a produtividade média é incrementada 9.364 unidades, em condições constantes das restantes variáveis.

A construção do modelo final, que permitiu relacionar os fatores supracitados com a produtividade das parcelas, consistiu num processo pautado de dificuldades a diversos níveis. Por um lado, salienta-se a dificuldade inerente à recolha dos dados da temperatura, visto que não foi possível a obtenção de dados através do IPMA (Instituto Potuguês do Mar e da Atmosfera), por falta de resposta desta entidade. Também o pedido de dados efetuado à Universidade de Aveiro não foi profícuo na medida em que os dados fornecidos eram relativos apenas ao distrito de Aveiro. Com efeito, por forma a serem obtidos os dados referentes à temperatura, face à sua importância na presente investigação, foi necessário proceder à sua recolha manual a partir do site *worldweatheronline*. Outra dificuldade relaciona-se com as características inerentes às variáveis do conjunto de dados fornecido, nomeadamente a existência de variáveis *dummy* (ou categóricas), o que implicou a obtenção de valores de $R^2$ reduzidos, característicos de regressão linear nas referidas condições. Salienta-se, por fim, a dificuldade sentida na construção de um modelo de regressão linear onde os pressupostos inerentes aos resíduos do mesmo, não são validados. Com efeito, foi necessário construir diversos modelos, por forma a averiguar uma solução profícua na validação dos pressupostos dos resíduos, ou consistente na sua presença. A solução encontrada nesse sentido foi a aplicação de modelos inovadores, e cujo desenvolvimento é ainda uma novidade.

Durante o estágio curricular foi possível efetuar uma visita ao campo, onde foi possível tomar consciência do procedimento inerente à classificação do nível de ataque. Com efeito, verificou-se que a monitorização é feita por 2 técnicos, em épocas distintas da realização do inventário florestal. Tal procedimento acarreta anomalias no estudo e na deteção da espécie invasora, visto que o AMA_U_Proj12 aquando do inventário florestal pode corresponder a valores característicos da inexistência de praga na parcela. Não obstante, caso a monitorização da praga seja realizada meses após o inventário, pode suceder que a presença da mesma tenha implicado a redução do AMA_U_Proj12, pelo que as conclusões podem ser inviabilizadas. Assim, como sugestão de melhoria, propõem-se a realização do inventário florestal e da monitorização da praga simultaneamente por forma a tornar os resultados e conclusões inerentes aos estudos mais realistas. Nesse sentido, a monitorização da praga deveria ser implementada continuamente em todas as parcelas por forma a desenvolver medidas de prevenção e de atuação na praga mais profícuas minimizando assim as consequências inerentes à sua existência.

Evidencia-se ainda, a necessidade de desenvolver a monitorização da praga de uma forma mais contínua e cuidada, em parcelas localizadas junto a linhas/cursos de água. Tal justifica-se pela incapacidade de desenvolver um combate da praga eficaz nestas zonas, visto que o controlo do *Gonipterus platensis*, em termos químicos, não é possível, tendo em conta que o uso de produtos químicos em linhas de água é punido por lei, pelo que, na presença de praga, o seu combate nestas zonas não se mostra tão profícuo pois o combate biológico nem sempre se mostra frutífero. Desta forma, uma monitorização deficitária nestas zonas pode implicar a sua

rápida propagação podendo, em casos extremos, implicar mesmo consequências desastrosas e irreversíveis, ao nível da perda de madeira. Desta forma, sugere-se à entidade de acolhimento uma monitorização mais cuidada e atenta nas parcelas localizadas nestas zonas por forma a evitar a existência de pragas nessas parcelas, tendo em conta a dificuldade inerente ao seu combate.

No que concerne à monitorização da praga, salienta-se o seu cariz subjetivo, visto que esta é realizada tendo por base os conhecimentos e inferência dos 2 técnicos responsáveis pelo referido procedimento. Desta forma, caso a monitorização seja implementada por diferentes técnicos, sendo pautada de elevada subjetividade, pode implicar diferentes classificações da praga.

No que concerne à dificuldade sentida na recolha das temperaturas, por forma a implementar investigações mais precisas, sugere-se à entidade de acolhimento a sua medição, e se possível, que esta seja realizada de uma forma rigorosa em cada parcela monitorizada. Tal justifica-se tendo em conta a sua importância na influência da produtividade e na atuação da espécie invasora, pelo que considerá-la na investigação no referido âmbito considera-se como primordial.

Por fim, salienta-se as consequências desastrosas provocadas pela atuação do *Gonipterus platensis*, em especial o ataque muito forte, provocando perdas de madeira totais, o que, a nível económico provoca um impacto negativo muito forte. Com efeito, a monitorização e a prevenção adequadas apresentam-se como primordiais e essenciais para garantir a minimização da perda de madeira, e por conseguinte os efeitos económicos negativos inerentes à atuação da espécie invasora, principalmente na indústria papeleira.

# Bibliografia


Alin, A. (2010). Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(3), 370–374.

Arshad, M., Rasool, M., & Ahmad, M. (2003). Anderson darling and modified anderson darling tests for generalized pareto distribution. *Pakistan Journal of Applied Sciences*, *3*(2), 85–88.

Associação da Indústria Papeleira – CELPA e Instituto Nacional de Investigação Agrária e Veterinária. (2015). Manual de boas práticas gorgulho-do-eucalipto [Computer software manual].

Barnett, V., & Lewis, T. (1974). *Outliers in statistical data*. John Wiley & Sons.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 211–252.

Branco, J., & Pires, A. (2007). Introdução aos métodos estatísticos robustos. *Edições SPE*.

Branco, M., Grodzi, W., Jacquet, J.-S., Moreira, F., Netherer, S., Schelhaas, M.-J., & Tomé, M. (2011). Report on specific risk analysis in regional forests of europe under various forest management alternatives. *Report on specific risk analysis in regional forests of Europe under various Forest Management Alternatives, EFI Technical Report 67, European Forest Institute*, 19–25.

Brown, P. J. (1993). *Measurement, regression and calibration*. Oxford: Clarendon Press.

Chatterjee, S., & Hadi, A. S. (2012). *Regression analysis by example* (5th ed.). John Wiley & Sons.

Cluster Habitat Sustentável. (2018). *Raiz - instituto de investigação da floresta e papel*. Disponível em `http://www.centrohabitat.net/pt/associado/raiz-instituto`

-de-investigacao-da-floresta-e-papel

Conover, W. J. (1999). Several $k$-sample kolmogorov-smirnov tests. *Ann. Math. Statist.*, *36*(3), 1019–1026. Disponível em https://doi.org/10.1214/aoms/1177700073

Dufour, J.-M., Farhat, A., Gardiol, L., & Khalaf, L. (1998). Simulation-based finite sample normality tests in linear regressions. *The Econometrics Journal*, *1*(1), 154–173.

Dufour, J.-M., Khalaf, L., Bernard, J.-T., & Genest, I. (2004). Simulation-based finite-sample tests for heteroskedasticity and arch effects. *Journal of Econometrics*, *122*(2), 317–347.

Grupo Portucel Soporcel. (2006). *Investigação nas áreas da floresta e do papel. uma renovação de raiz.* BCSD Portugal.

Guimarães, R., & Cabral, J. (2010). *Estatística* (2nd ed.). Lisboa, Portugal: Verlag Dashöfer.

Guo, S., Zhong, S., & Zhang, A. (2013). Privacy-preserving kruskal–wallis test. *Computer methods and programs in biomedicine*, *112*(1), 135–145.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2010). *Multivariate data analysis.* Prentice Hall.

Hall, A., Pereira, A. F., & Neves, C. (2011). *Grande maratona de estatística no spss.* Lisboa: Escolar Editora.

Hamburg, M., & Young, P. (1994). *Statistical analysis for decision making.* Duxbury Press.

Heritier, S., Cantoni, E., Copt, S., & Victoria-Feser, M.-P. (2009). *Robust methods in biostatistics* (Vol. 825). John Wiley & Sons.

Instituto da Conservação da Natureza e das Florestas, Direção-Geral de Alimentação e Veterinária, Instituto Nacional de Investigação Agrária e Veterinária, Associação da Indústria Papeleira – CELPA, Grupo Portucel Soporcel, Instituto de Investigação da Floresta, Papel – RAIZ e Altri florestal. (2015). Plano de controlo para o inseto *Gonipterus platensis* [Computer software manual].

Johnson, R. A. (81). *Wichern, dw (1998), applied multivariate statistical analysis* (Vol. 7632).

Krämer, W., & Sonnberger, H. (1986). Testing disturbances. In *The linear regression model under test* (pp. 17–42). Heidelberg: Physica-Verlag HD.

Krzanowski, W. (1995). *Recent advances in descriptive multivariate analysis.* Oxford: Clarendon Press.

Leotti, V. B., Birck, A. R., & Riboldi, J. (2005). Comparação dos testes de aderência à normalidade kolmogorov-smirnov, anderson-darling, cramer–von mises e shapiro-wilk por

simulação. *Anais do 11º Simpósio de Estatística Aplicada à Experimentação Agronômica*.

Leotti, V. B., Coster, R., & Riboldi, J. (2012). Normalidade de variáveis: métodos de verificação e comparação de alguns testes não-paramétricos por simulação. *Revista HCPA*, *32*(2), 227–234.

Lewis-Beck, M. S. (1993). *Regression analysis*. London: Sage.

Li, P. (2005). Box-cox transformations: an overview.

Lim, T.-S., & Loh, W.-Y. (1996). A comparison of tests of equality of variances. *Computational Statistics & Data Analysis*, *22*(3), 287–301.

Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297–303.

Mapondera, T. S., Burgess, T., Matsuki, M., & Oberprieler, R. G. (2012). Identification and molecular phylogenetics of the cryptic species of the gonipterus scutellatus complex (coleoptera: Curculionidae: Gonipterini). *Austral Entomology*, *51*(3), 175–188.

Marôco, J. (2010). *Análise estatística com o pasw statistics* (3a Edição ed.). Pêro Pinheiro: Report Number.

Mardia, K., Kent, J., & Bibby, J. (1994). *Multivariate analysis*. London: Academic Press.

Maronna, R. A. (2011). Robust ridge regression for high-dimensional data. *Technometrics*, *53*(1), 44–53.

McLeod, A. I., & Li, W. K. (1983). Diagnostic checking arma time series models using squared-residual autocorrelations. *Journal of Time Series Analysis*, *4*(4), 269–273.

Miles, J. (2005). Tolerance and variance inflation factor. *Encyclopedia of statistics in behavioral science*.

Murteira, B., Ribeiro, C. S., e Silva, J. A., & Pimenta, C. (2010). *Introdução à estatística*. Lisboa: Escolar Editora.

Oliveira, M. M., Santos, L. D., & Fortuna, N. (2011). *Econometria*. Lisboa: Escolar Editora.

Peña, E. A., & Slate, E. H. (2006). Global validation of linear model assumptions. *Journal of the American Statistical Association*, *101*(473), 341–354.

Rahman, M. M., & Govindarajulu, Z. (1997). A modification of the test of shapiro and wilk for normality. *Journal of Applied Statistics*, *24*(2), 219–236.

RAIZ. (s.d.). *Raiz - instituto de investigação da floresta e do papel | raiz*. Disponível em

http://raiz-iifp.pt/{#}

Rana, M. S., Midi, H., & Imon, A. R. (2008). A robust modification of the goldfeld-quandt test for the detection of heteroscedasticity in the presence of outliers. *Journal of mathematics and Statistics*, *4*(4), 277.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, *2*(1), 21–33.

Reboredo, F. (2014). *Forest context and policies in portugal: Present and future challenges* (Vol. 19). Springer.

Reis, A. R., Ferreira, L., Tomé, M., Araujo, C., & Branco, M. (2012). Efficiency of biological control of gonipterus platensis (coleoptera: Curculionidae) by anaphes nitens (hymenoptera: Mymaridae) in cold areas of the iberian peninsula: implications for defoliation and wood production in eucalyptus globulus. *Forest Ecology and management*, *270*, 216–222.

Ryan, T. P. (2008). *Modern regression methods* (Vol. 655). John Wiley & Sons.

Sakia, R. (1992). The box-cox transformation technique: a review. *The statistician*, 169–178.

Stapleton, J. H. (2009). *Linear statistical models* (Vol. 719). John Wiley & Sons.

The Navigator Company. (s.d.). *Investigação e Desenvolvimento | The Navigator Company.* Disponível em http://www.thenavigatorcompany.com/Pasta-e-Papel/Investigacao-e-Desenvolvimento

The Navigator Company. (2016). *Relatório e contas 2016.* Disponível em http://www.thenavigatorcompany.com/var/ezdemo_site/storage/original/application/f7779c2a6231626b2b6b25586d79f4fc.pdf

Tomé, M., Oliveira, T., & Soares, P. (2006). O modelo globulus 3.0-dados e equações. *Publicações do GIMREF*, 1–26.

Torrão, S. (2017). *Os eucaliptos e as aves da quinta de são francisco.* Lisboa: Medialivros – Actividades Editoriais, S.A.

Valente, C., Vaz, A., Pina, J., Manta, A., & Sequeira, A. (2004). Control strategy against the eucalyptus snout beetle, gonipterus scutellatus gyllenhal (coleoptera, curculionidae), by the portuguese cellulose industry. In *Eucalyptus in a changing world. procedings of iufro conference, aveiro* (pp. 622–627).

Verbeek, M. (2004). *A guide to modern econometrics*. John Wiley & Sons.

Weisberg, S. (2001). Yeo-johnson power transformations. *Department of Applied Statistics, University of Minnesota. Retrieved June*, *1*, 2003.

Wikipédia. (2008). *Portucel – wikipédia, a enciclopédia livre.* Disponível em `https://pt.wikipedia.org/wiki/Portucel`

Yamashita, T., Yamashita, K., & Kamimura, R. (2007). A stepwise aic method for variable selection in linear regression. *Communications in Statistics - Theory and Methods*, *36*(13), 2395-2403.

Yeo, I.-K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954–959.

Zeileis, A. (2004). Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, *11*, 1–17.

# Apêndice A

# Descrição das variáveis do conjunto de dados

Tabela A.1: Variáveis do conjunto de dados

| Variável | Designação | Unidade |
|---|---|---|
| **xcoord** | Coordenada geográfica do centro da parcela (abcissa) | |
| **ycoord** | Coordenada geográfica do centro da parcela (ordenada) | |
| **Id** | Designação da parcela através do seu número de identificação | |
| **fid** | Confirmação da variável Id | |
| **ID UG** | Identificação da área de estudo | |
| **UG_OCUP** | Unidade de gestão por talhão | |
| **Talhão** | Identificação do talhão | |
| **TALHAO_OCUP** | Identificação de talhão e ocupação | |
| **Ano** | Ano que foi feito o inventário | |
| **Parcela** | Designação da parcela | |
| **Data Real Medição** | Data real da medição | |
| **Hdom** | Altura dominante, média das alturas das 100 árvores mais grossas por ha (árvores dominantes) | m |
| **Ddom** | Diâmetro dominante, média dos diâmetros das 100 árvores mais grossas por ha (árvores dominantes) | m |

| | | |
|---|---|---|
| **Npl** | Número de árvores plantadas | $ha^{-1}$ |
| **N** | Número de árvores plantadas na parcela | $ha^{-1}$ |
| **N_vivas** | Número de árvores vivas; na talhadia, número de varas vivas | |
| **Sob_toicas** | Percentagem de toicas vivas – 1ª rotação | |
| **Sob_rot** | Percentagem de toicas vivas – talhadia | |
| **N_varas** | Número total de varas | |
| **N_varas_vivas** | Número de varas vivas | |
| **N_varas_toica** | Número de varas numa toiça | |
| **N_varas_vivas_toica** | Número de varas vivas numa toiça, ou seja, número de varas vivas que sejam altas e grossas | |
| **Sob_varas** | Percentagem de varas vivas | |
| **Area_arv** | Área ocupada por uma árvore | $m^2$ |
| **G** | Área basal | $m^2/ha$ |
| **S** | Índice de qualidade da estação (site *index*) | |
| **Dg** | Diâmetro da árvore de área seccional média do povoamento | m |
| **Rot** | 1ª rotação (plantação); 2ª rotação (rebrota a partir de toiça) | |
| **ROT_OCUP** | 1ª rotação (plantação); 2ª rotação (rebrota a partir de toiça) - Confirmação valores de ROT | |
| **PC_IDADE** | Idade da parcela | Anos |
| **CLASSE_IDADE** | Classe de idade | Anos |
| **dif_idade** | Diferença de idades (confirmação apenas) | Anos |
| **Regiao_Glob** | Classificação da região a partir de *globulus* | |
| **MG** | Material genético | |
| **MG_OCUP** | Material genético (clonal – EG_matGenet e seminal – EG) | |
| **Clima** | Região climática (1 a 10 – 1 reduzida aptidão, 10 de elevada aptidão) | |
| **Solo** | Região ou classe de solo (1 a 10 – 1 reduzida aptidão, 10 de elevada aptidão) | |
| **RP_Glob** | Região de produtividade (1 a 8 – 1 elevada produtividade, 8 reduzida produtividade) | |
| **REGIAO** | Região 1 (RP1, 2 e 3); Região 2 (RP 4, 5 e 6); Região 3 (RP 7 e 8) | |
| **dias_pp** | Dias de precipitação | |
| **cota** | Altitude | m |
| **PC_AREA_AJUST** | Área da parcela ajustada | $m^2$ |
| **PC_DECLIV** | Declive da parcela | |
| **dt_ref_OC** | Data de plantação ou corte | |
| **PREP_TERR** | Tipo de preparação do terreno (terraços, vala comoro e sem armação) | |

| | | |
|---|---|---|
| **V_mt** | Volume com casca e com cepo, volume total estimado com equações de Margarida Tomé (MT) | $m^3$ |
| **Vu_mt** | Volume sem casca e com cepo estimado com equações de MT | $m^3$ |
| **Vb** | Volume da casca | $m^3$ |
| **V_st_mt** | Volume com casca e sem cepo estimado com equações de MT | $m^3$ |
| **Vst** | Volume do cepo com casca | $m^3$ |
| **Vu_st_mt** | Volume sem casca e sem cepo estimado com equações de MT | $m^3$ |
| **Vmdi_6_5_mt** | Volume mercantil com casca (sem cepo) para um diâmetro de desponta (di) de 6.5 cm estimado com equações de MT | $m^3$ |
| **Vmudi_5_5_mt** | Volume mercantil sem casca (sem cepo) para um diâmetro de desponta sem casca (di) de 6.5 cm estimado com equações de MT | $m^3$ |
| **AMA** | Acréscimo médio anual | $m^3$/ha/ano |
| **AMA_U** | Acréscimo médio anual útil | $m^3$/ha/ano |
| **Hdom_Proj12** | Altura dominante, média das alturas das 100 árvores mais grossas por ha (árvores dominantes), projetada a 12 anos | m |
| **N_Proj12** | Número de árvores da parcela, projetado a 12 anos | |
| **G_Proj12** | Área basal da árvore projetada a 12 anos | $m^2$ |
| **Vu_Proj12** | Volume sem casca e com cepo projetado a 12 anos | $m^3$ |
| **Vb_Proj12** | Volume da casca projetado a 12 anos | $m^3$ |
| **Vst_Proj12** | Volume do cepo com casca projetado a 12 anos | $m^3$ |
| **V_Proj12** | Volume com casca e com cepo, volume total projetado a 12 anos | $m^3$ |
| **AMA_Proj12** | Acréscimo médio anual útil calculado com volumes projetados a 12 anos | $m^3$/ha/ano |
| **Vmudi_55_Proj12** | Volume mercantil sem casca (sem cepo) para um diâmetro de desponta sem casca (di) de 6.5 cm projetado a 12 anos | $m^3$ |
| **AMA_U_Proj12** | Acréscimo médio anual útil calculado com volumes úteis projetados a 12 anos | $m^3$/ha/ano |
| **cod_ug** | Confirmação da variável UG_OCUP | |
| **cod_talhao** | Confirmação da variável TALHAO_OCUP | |
| **rotacao** | | |
| **MG_1** | Confirmação da variável MG | |
| **n_ataque** | Nível de ataque do *Gonipterus platensis* com classificação numérica (0 a 5) | |
| **obs** | Observações das parcelas com classificação numérica (0 a 8) | |

| | | |
|---|---|---|
| **inseticida** | Aplicação de inseticida (0 – não aplicado, 1 – Aplicado) | |
| **psila** | Presença da praga Psila (0 – Ausência, 1 – presença) | |
| **ataque** | Nível de ataque do *Gonipterus platensis* com classificação descritiva | |
| **Observação** | Observações das parcelas com classificação descritiva | |
| **Psila_1** | Relevância do ataque da psila (relevante e não relevante) | |
| **id_1** | Confirmação da variável Id | |
| **dicomun** | Designação comum da área de estudo | |
| **Distrito** | Distrito onde se encontra a parcela | |
| **Municipios** | Município onde se encontra a parcela | |
| **taa** | Identificação do tipo de área administrativa | |
| **area_ea_ha** | Valor área da área administrativa | ha |
| **area_t_ha** | Valor total da área dos municípios | ha |

# Apêndice B

# Análises das variáveis do conjunto de dados

## B.1 Parcelas avaliadas/não avaliadas

Percentagem de parcelas com e sem informação de avaliação do *Gonipterus platensis* no conjunto dos 5 anos.

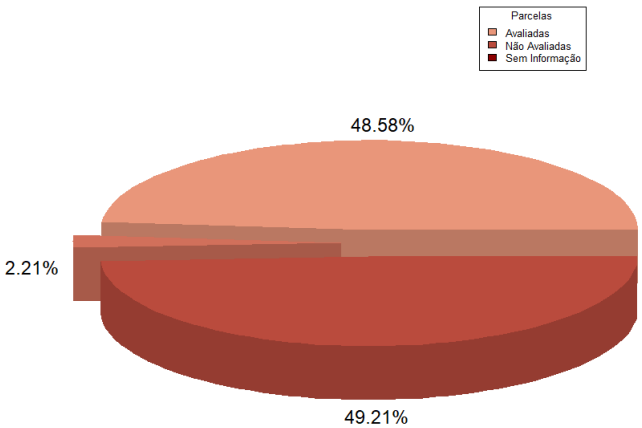



Figura B.1: Percentagem de parcelas com e sem informação de avaliação do *Gonipterus platensis* no conjunto dos 5 anos

Há uma enorme percentagem, quase metade das parcelas que não contêm qualquer in-

formação, isto pode ser pelo facto de se encontrarem no sul de Portugal e a presença do *Gonipterus platensis* ainda não ter sido notada. Como podíamos ver no mapa anterior a informação sobre a praga encontra-se do centro de Portugal para Norte.

Relativamente às avaliadas e às que contêm informação sobre não avaliadas podemos ver na tabela seguinte.

Tabela B.1: Número de parcelas avaliadas e classificadas como não avaliadas

| | Ano 2011 | Ano 2012 | Ano 2013 | Ano 2014 | Ano 2015 |
|---|---|---|---|---|---|
| Avaliadas | 1600 | 2165 | 2624 | 2904 | 2805 |
| Não Avaliadas | 49 | 95 | 124 | 146 | 136 |

É de se fazer notar que o número de parcelas com o registo de não avaliadas tem vindo a aumentar assim como o registo de avaliadas. Veremos em termos de percentagens.

Tabela B.2: Percentagem de parcelas avaliadas e classificadas como não avaliadas

| | Ano 2011 | Ano 2012 | Ano 2013 | Ano 2014 | Ano 2015 |
|---|---|---|---|---|---|
| % Avaliadas | 97.03 | 95.8 | 95.49 | 95.21 | 95.38 |
| % Não Avaliadas | 2.97 | 4.2 | 4.51 | 4.79 | 4.62 |

Em termos absolutos vimos que o número de parcelas monitorizadas tem vindo a aumentar, em comparação em termos percentuais das avaliadas pelas não avaliadas vemos que o número não tem sofrido grandes alterações, dentro deste grupo as parcelas avaliadas rondam os 95%. É de notar que o número de parcelas monitorizadas tem vindo a aumentar de ano para ano.

Tabela B.3: Observações por nível de ataque

| | 0 – Não Avaliado | 1 – Inexistente | 2 – Fraco | 3 – Moderado | 4 – Forte | 5 – Muito Forte |
|---|---|---|---|---|---|---|
| A corte ou cortado | 34 | 26 | 7 | 23 | 0 | 0 |
| Abandonado ou contrato de neg. | 21 | 0 | 0 | 0 | 0 | 0 |
| Ardido | 24 | 0 | 0 | 0 | 0 | 0 |
| Doenças no tronco | 0 | 0 | 55 | 0 | 0 | 0 |
| Geada | 0 | 0 | 3 | 5 | 0 | 0 |
| Inferido | 13 | 3604 | 47 | 46 | 8 | 3 |
| Outra espécie de eucalipto | 36 | 0 | 0 | 0 | 0 | 0 |
| Outros | 76 | 0 | 0 | 0 | 0 | 0 |
| Sem observações | 0 | 2437 | 2973 | 2018 | 679 | 166 |

## B.2 Contabilização do nível de ataque por ano

Tabela B.4: Contabilização do nível de ataque por ano

| | Ano 2011 | Ano 2012 | Ano 2013 | Ano 2014 | Ano 2015 |
|---|---|---|---|---|---|
| 1 – Inexistente | 751 | 1179 | 1313 | 1382 | 1440 |
| 2 – Fraco | 432 | 479 | 702 | 769 | 703 |
| 3 – Moderado | 286 | 362 | 465 | 511 | 4682 |
| 4 – Forte | 105 | 116 | 125 | 194 | 147 |
| 5 – Muito Forte | 26 | 29 | 19 | 48 | 47 |

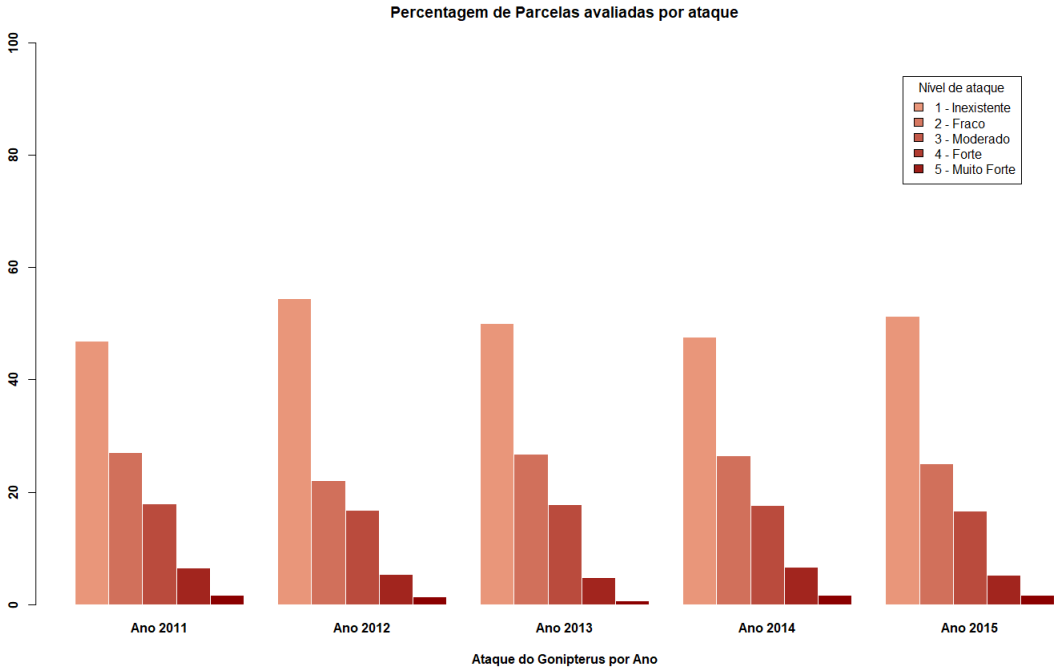



Figura B.2: Percentagem do nível de ataque por ano

## B.3 Classe clima por nível de ataque

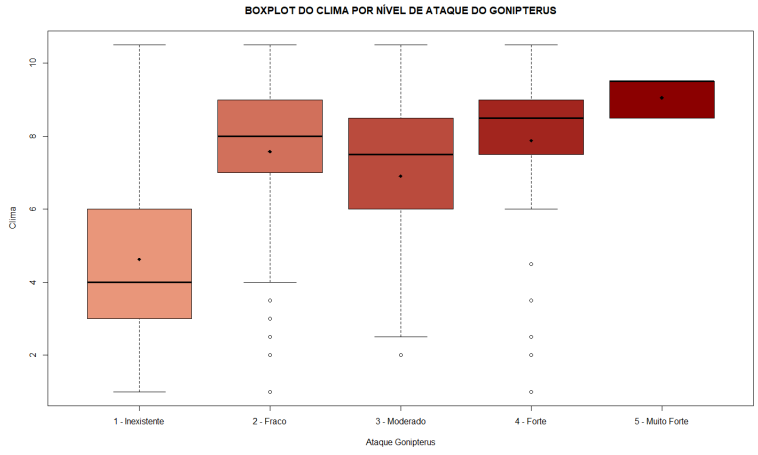



Figura B.3: Caixa de bigodes clima por ataque do *Gonipterus platensis*

Na caixa de bigodes apresentada anteriormente, referente à classe clima por nível de ataque da praga, observa-se a existência de 415 *outliers* que se encontram distribuídos pelos níveis de ataque fraco, moderado e forte, correspondendo, respetivamente, aos valores 246, 93 e 76.

## B.4 AMA Útil Proj12 por ano

A conclusão de que o AMA_U_Proj12 diminui à medida que o nível de ataque é cada vez mais forte, ou seja, à medida que a presença do gorgulho do eucalipto é cada vez mais evidente, é também comprovada com base na representação da figura B.4. Com efeito, a sua análise permite evidenciar que, para cada ano em estudo, o AMA_U_Proj12 médio diminui com a intensificação do nível de ataque presente nas parcelas em estudo. Outra observação que é possível fazer-se é que o comportamento do AMA_U_Proj12 médio em função do nível de ataque através do qual as parcelas são classificadas é bastante semelhante ao longo dos anos. Os valores mais elevados de AMA_U_Proj12 médio são verificados em parcelas cujo nível de ataque é inexistente ou fraco, sendo mais reduzidos em parcelas cujo nível de ataque do *Gonipterus platensis* se classifica como sendo forte ou muito forte.

Ainda relativamente às caixas de bigodes apresentadas, observa-se a existência de valores *outliers*, visíveis pelos pontos localizados acima das caixas de bigodes. Com efeito, foi detetada a existência de 280 valores *outliers*, dos quais 133 são parcelas classificadas com nível de ataque inexistente, 81 são classificadas com nível de ataque fraco, 46 classificam-se com ataque moderado, 18 cujo nível de ataque é forte e 2 com nível de ataque muito forte. Os *outliers* detetados representam cerca de 2% da quantidade de dados em estudo, pelo que o número diminuto da quantidade de *outliers* justifica a sua não remoção.

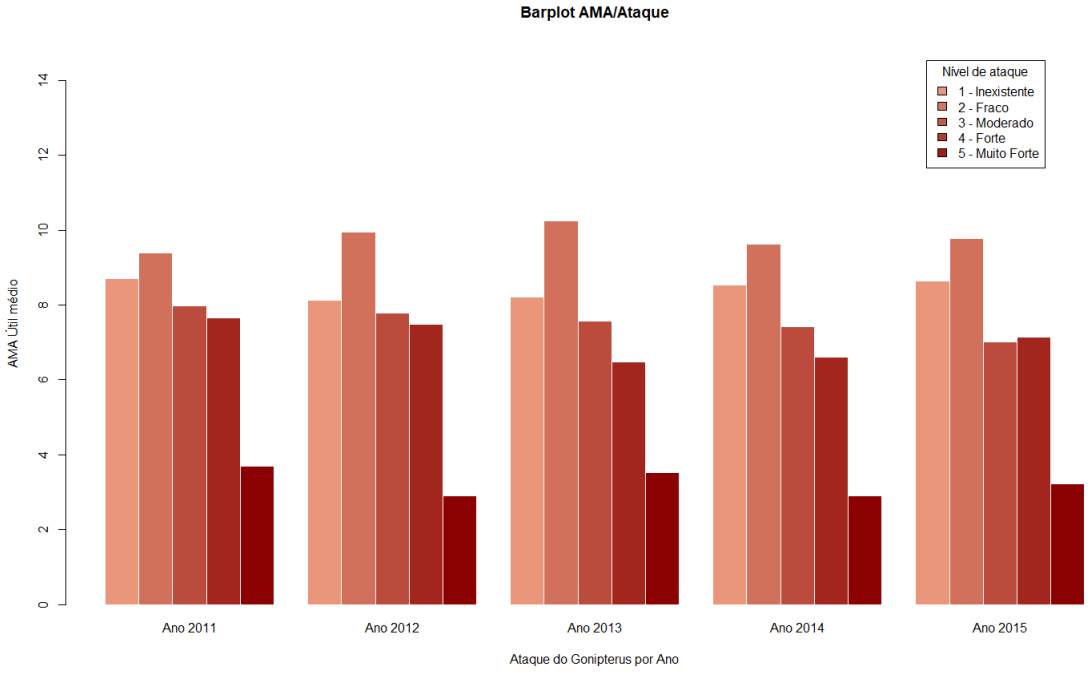



Figura B.4: Representação do AMA Útil Médio por nível de ataque

A análise do gráfico de barras apresentado na figura B.4 permite ainda supor que o valor do AMA_U_Proj12 médio não tenha sofrido muitas alterações ao longo dos anos em estudo,

o que se pode comprovar com base nas caixas de bigodes que relacionam, para cada ano em estudo, a caixa de bigodes referente ao AMA_U_Proj12 médio. Com efeito, verifica-se que as caixas de bigodes são semelhantes ao longo dos anos em estudo, os valores médios e medianos de AMA_U_Proj12 médio ao longo dos anos, são próximos entre si. O que pode ser comprovado com a aplicação do teste de Kruskal-Wallis que permite a obtenção de um valor de p-value de 0.2795, maior que o nível de significância de 0.05, o que permite concluir que não existem diferenças significativas entre as caixas de bigodes.

Tabela B.5: Estatísticas sumárias do AMA_U_Proj12 por ano

| group | count | mean | sd | median | IQR |
|---|---|---|---|---|---|
| <ord> | <int> | <dbl> | <dbl> | <dbl> | <dbl> |
| 2011 | 1600 | 8.61 | 5.29 | 7.58 | 6.62 |
| 2012 | 2165 | 8.37 | 5.35 | 7.32 | 6.57 |
| 2013 | 2622 | 8.53 | 5.22 | 7.53 | 6.59 |
| 2014 | 2904 | 8.42 | 5.31 | 7.41 | 6.72 |
| 2015 | 2805 | 8.49 | 5.36 | 7.50 | 6.77 |

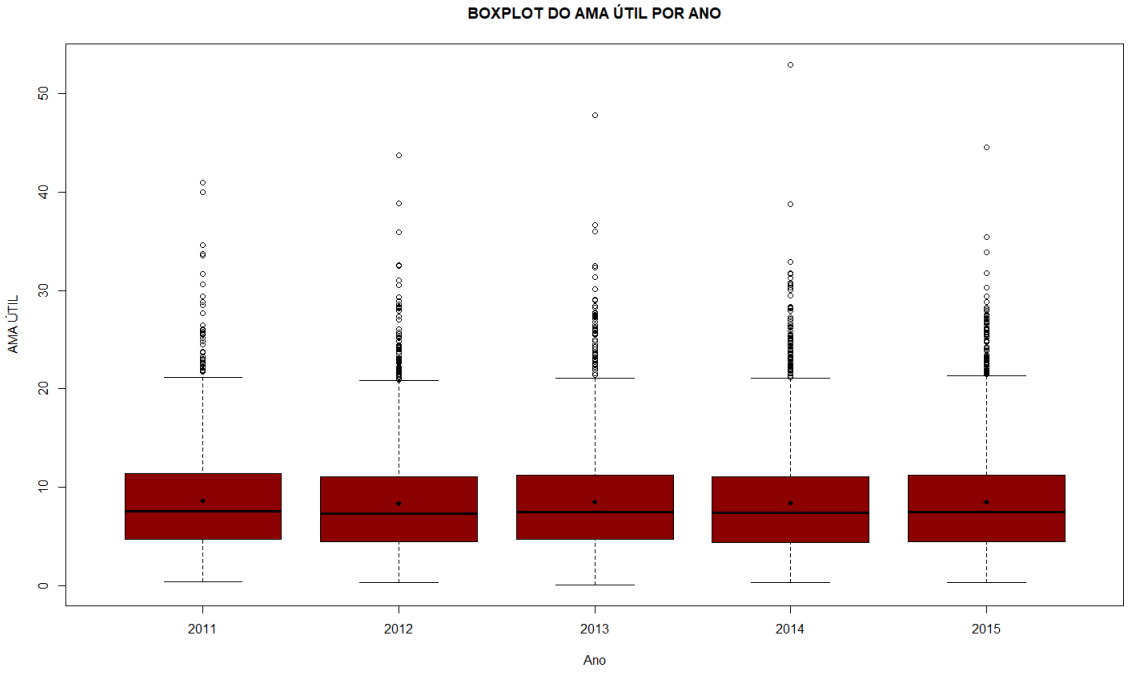



Figura B.5: Caixa de bigodes relativo ao AMA por ano

A análise gráfica das caixas de bigodes apresentadas permite detetar a existência de *outliers* para os cinco anos em estudo, num total de 319 *outliers*, distribuídos da forma: 38 *outliers* em 2011, 64 *outliers* em 2012, 61 *outliers* em 2013, 80 *outliers* em 2014 e 76 *outliers* em 2015.

# Apêndice C

# Modelo de regressão linear múltipla

## C.1 Modelo 1

O presente modelo, construído utilizando o AMA_U_Proj12 como variável dependente e todas as variáveis explanadas anteriormente como variáveis independentes, permitiu alcançar os seguintes resultados.

Tabela C.1: Resultados da aplicação da regressão OLS ao modelo 1

| | *Variável dependente:* |
|---|---|
| | AMA_U_Proj12 |
| n_ataque 2 | −0.900*** |
| | (0.118) |
| n_ataque 3 | −2.544*** |
| | (0.133) |
| n_ataque 4 | −3.604*** |
| | (0.190) |
| n_ataque 5 | −9.348*** |
| | (0.351) |
| TMinF | 0.203*** |
| | (0.037) |
| TMaxF | −0.018 |
| | (0.041) |
| TMinQ | −0.089** |
| | (0.032) |
| TMaxQ | −0.114*** |
| | (0.032) |
| dias_pp | 0.074*** |
| | (0.002) |
| N_fr | 9.911*** |
| | (0.200) |
| cota | 0.001** |
| | (0.0003) |
| Constant | −0.438 |
| | (0.905) |
| Observations | 12,096 |
| $R^2$ | 0.366 |
| Adjusted $R^2$ | 0.365 |
| Residual Std. Error | 4.226 (df = 12082) |
| F Statistic | 634.1*** (df = 11; 12084) |
| Nota: | *p<0.1; **p<0.05; ***p<0.01 |

Verificou-se que a aplicação da metodologia de seleção de variáveis *Stepwise* promoveu a remoção das variável e TMaxF na tabela C.2.

Tabela C.2: Resultados do *Stepwise* no modelo 1

| | Df | Sum of Sq | RSS | AIC |
|---|---|---|---|---|
| <none> | | | 215786 | 34876 |
| + TMaxF | 1 | 3 | 215783 | 34877 |
| − TMinQ | 1 | 135 | 215921 | 34881 |
| − cota | 1 | 160 | 215947 | 34883 |
| − TMaxQ | 1 | 599 | 216385 | 34907 |
| − TMinF | 1 | 725 | 216511 | 34914 |
| − dias_pp | 1 | 16526 | 232312 | 35766 |
| − n_ataque | 4 | 19069 | 234855 | 35892 |
| − N_fr | 1 | 44404 | 260190 | 37137 |

*Note:* (+) Remoção da variável; (−) Manter variável

Os resultados da regressão linear OLS do modelo 1 com a aplicação de *Stepwise* estão sintetizados na tabela C.3.

Tabela C.3: Resultados da aplicação do *Stepwise* na regressão OLS ao modelo 1

| | *Variável dependente:* |
|---|---|
| | AMA_U_Proj12 |
| n_ataque 2 | −0.892*** |
| | (0.117) |
| n_ataque 3 | −2.538*** |
| | (0.132) |
| n_ataque 4 | −3.601*** |
| | (0.190) |
| n_ataque 5 | −9.342*** |
| | (0.351) |
| TMinF | 0.193*** |
| | (0.030) |
| TMinQ | −0.085*** |
| | (0.031) |
| TMaxQ | −0.116*** |
| | (0.0200) |
| N_fr | 9.899*** |
| | (0.199) |
| cota | 0.001** |
| | (0.0003) |
| dias_pp | 0.074*** |
| | (0.002) |
| Constant | −0.637 |
| | (0.782) |
| Observations | 12,096 |
| $R^2$ | 0.366 |
| Adjusted $R^2$ | 0.365 |
| Residual Std. Error | 4.226 (df = 12085) |
| F Statistic | 697.5*** (df = 10; 12085) |
| Nota: | *p<0.1; **p<0.05; ***p<0.01 |

Uma análise interpretativa dos resultados obtidos, permitiu indagar que, de acordo com o teste de significância de cada coeficiente, os valores do *p-value* são reduzidos (e, consequentemente, os coeficientes significativamente diferentes de zero com base num nível de significância de 5%). Tal característica é inerente a todos os coeficientes obtidos, à exceção de um: o valor

do *p-value* do teste de significância do coeficiente de referência (inexistência de ataque) não é significativo (o $p-value$ do *intercept* é 0.416), o que permite intuir que a ausência de ataque do *Gonipterus platensis* não influência diretamente o AMA_U_Proj12.

Com efeito, de uma forma geral, os coeficientes inerentes às variáveis consideradas são significativamente diferentes de zero.

Relativamente ao ajuste dos modelos de regressão linear, verifica-se que o valor do $R^2$ obtido foi de 0.3656.

Neste modelo o valor de estatística F obtido foi na ordem de 697.5, o que permite concluir que o modelo adequa-se aos dados.

A utilização da função *vif()* do *software* R às variáveis independentes do modelo de regressão linear construído, permitiu concluir a ausência de colinearidade entre as mesmas. Os resultados encontram-se sintetizados na tabela C.4.

Tabela C.4: Resultados da aplicação do *vif()* no modelo 1

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| N_fr | 1.109 | 1 | 1.053 |
| cota | 1.781 | 1 | 1.335 |
| dias_pp | 2.063 | 1 | 1.436 |
| n_ataque | 2.078 | 4 | 1.096 |
| TMinF | 2.274 | 1 | 1.508 |
| TMinQ | 1.920 | 1 | 1.386 |
| TMaxQ | 1.428 | 1 | 1.195 |

Os gráficos seguintes dizem respeito aos resíduos do modelo 1.

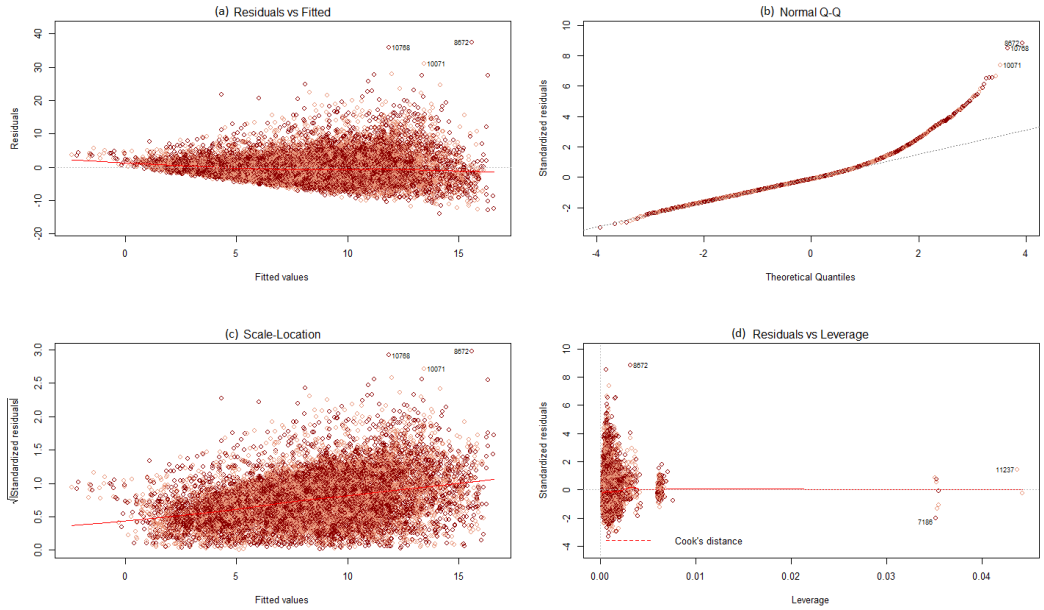



Figura C.1: Gráficos de resíduos do modelo 1 com regressão OLS

Pela a observação do gráfico e pela respetiva análise dos testes aos pressupostos da regressão linear, verifica-se a violação dos mesmo, desta forma aplicou-se a transformação de Box-Cox para tentar colmatar essas falhas.

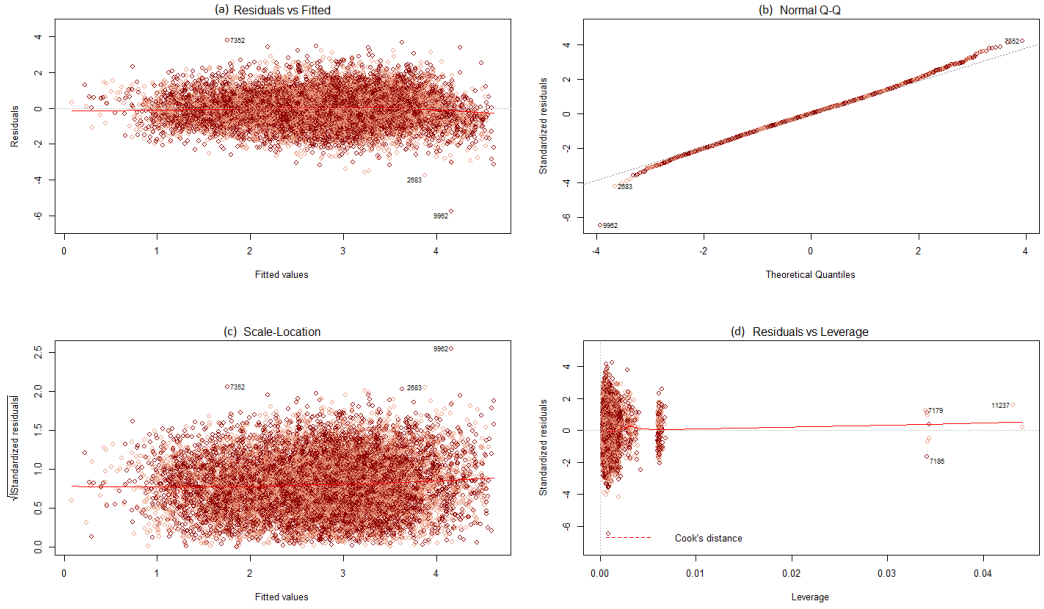



Figura C.2: Gráficos de resíduos do modelo 1 com transformação Box-Cox

## C.2 Métodos de estimação na presença de autocorrelação

A tabela C.5 foi conseguida através do *software E.Views*, por forma a aplicar o passo um do método de Durbin e assim encontrar o $\rho$ para que no passo 2 sejam encontrados os estimadores da regressão na presença de autocorrelação.

Tabela C.5: Resultados obtidos no passo 1 do método de Durbin

Dependent Variable: AMA_NEW3
Method: Least Squares
Sample (adjusted): 2 12096
Included observations: 12095 after adjustments

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | 0.363943 | 0.160921 | 2.261628 | 0.0237 |
| AMA_NEW3(-1) | 0.506413 | 0.007844 | 64.56380 | 0.0000 |
| TMINF | 0.010430 | 0.006560 | 1.589935 | 0.1119 |
| TMINF(-1) | 0.012026 | 0.006575 | 1.829197 | 0.0674 |
| TMAXQ | −0.006035 | 0.003827 | −1.576700 | 0.1149 |
| TMAXQ(-1) | −0.014457 | 0.003817 | −3.787709 | 0.0002 |
| N_FR | 2.514605 | 0.042806 | 58.74372 | 0.0000 |
| N_FR(-1) | −1.267470 | 0.047099 | −26.91096 | 0.0000 |
| DIAS_PP | 0.012162 | 0.001467 | 8.291141 | 0.0000 |
| DIAS_PP(-1) | −0.004154 | 0.001471 | −2.823737 | 0.0048 |
| N_ATAQUE=2 | −0.109201 | 0.042296 | −2.581849 | 0.0098 |
| N_ATAQUE=3 | −0.538837 | 0.048920 | −11.01457 | 0.0000 |
| N_ATAQUE=4 | −0.682624 | 0.065081 | −10.48879 | 0.0000 |
| N_ATAQUE=5 | −2.154681 | 0.150263 | −14.33938 | 0.0000 |
| N_ATAQUE(-1)=2 | 0.019181 | 0.042337 | 0.453060 | 0.6505 |
| N_ATAQUE(-1)=3 | 0.305837 | 0.049044 | 6.236010 | 0.0000 |
| N_ATAQUE(-1)=4 | 0.325445 | 0.065298 | 4.984024 | 0.0000 |
| N_ATAQUE(-1)=5 | 0.944181 | 0.151407 | 6.236041 | 0.0000 |

| | | | |
|---|---|---|---|
| R-squared | 0.579297 | Mean dependent var | 2.738221 |
| Adjusted R-squared | 0.578704 | S.D. dependent var | 1.185124 |
| S.E. of regression | 0.769232 | Akaike info criterion | 2.314639 |
| Sum squared resid | 7146.177 | Schwarz criterion | 2.325653 |
| Log likelihood | −13979.78 | Hannan-Quinn criter. | 2.318332 |
| F-statistic | 978.2167 | Durbin-Watson stat | 2.252449 |
| Prob(F-statistic) | 0.000000 | | |

## C.3 Deteção de *outliers*

Os gráficos das várias medidas de diagnóstico utilizadas para a deteção dos *outliers* podem ser consultados na figura C.3.

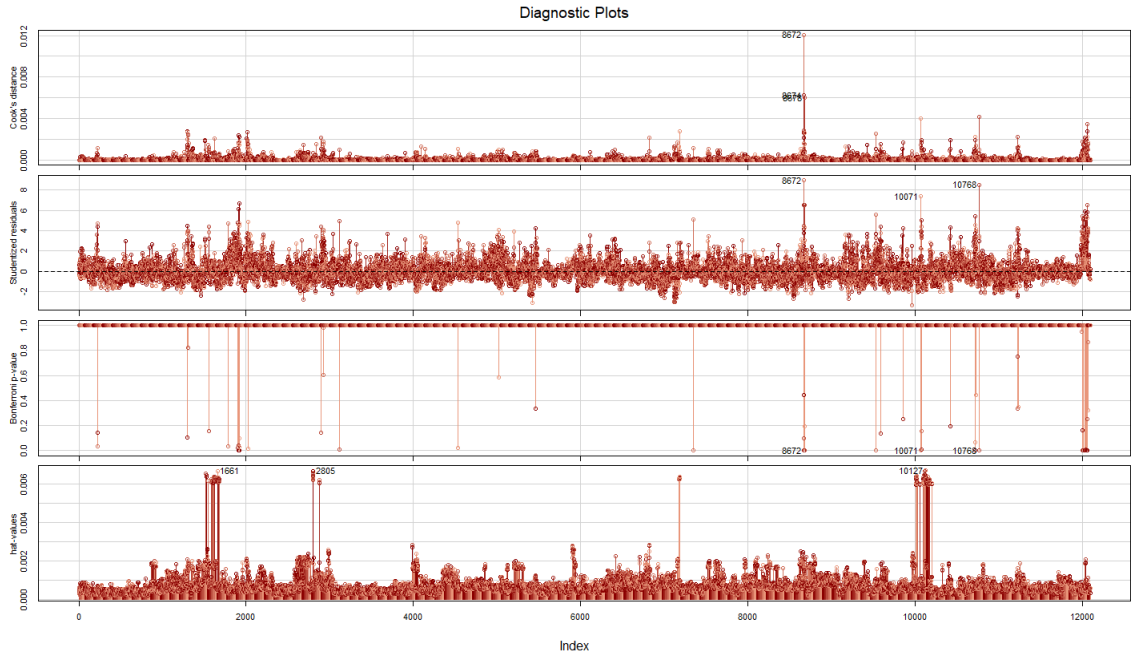



Figura C.3: Gráficos das medidas de diagnóstico dos *outliers*

É possível analisar o *leverage* para cada variável independente através dos seguintes gráficos.

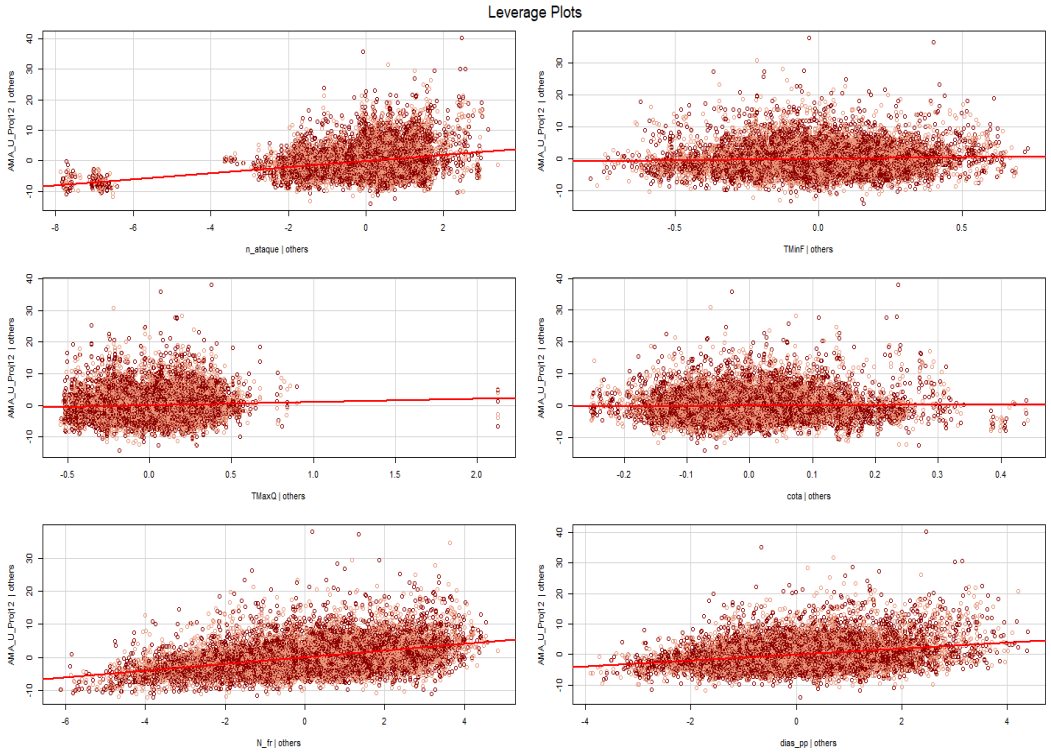



Figura C.4: Gráficos de *leverage*

# Apêndice D

# Código utilizado no *software* R

## D.1 Código utilizado na transformação Box-Cox

*Output* D.1: Input da Transformação Box-Cox

```
#Package
library(cars)

#Regressão OLS
model2 <- lm(AMA_U_Proj12 ~ as.factor(n_ataque) + TMinF + TMaxQ + cota
    + N_fr + dias_pp, data = DadosTotal)

#Transformação do AMA
distBCMod2 <- BoxCoxTrans(DadosTotal$AMA_U_Proj12)
print(distBCMod2)

#Adicionar a nova variável AMA_new ao conjunto de dados
DadosTotal <- cbind(DadosTotal, AMA_new3=predict(distBCMod3,
    DadosTotal$AMA_U_Proj12)) # append the transformed variable to
    DadosTotal

#Regressão OLS com a nova variável
lmMod_bc3 <- lm(AMA_new3 ~ N_fr + cota   + as.factor(n_ataque) + TMinF
```

```
    + TMaxQ + dias_pp, data = DadosTotal)
summary(lmMod_bc3)
```

## D.2 Código utilizado na HAC

*Output* D.2: Input da HAC

```
#Package
library(sandwich)

#Aplicação de HAC
cov <- vcovHAC(model2, type = "HAC")
robust.rse <- sqrt(diag(cov)) # robust standard errors
coeftest(model2, vcovHAC(model2 , type = "HAC")) # Heteroskedasticity
    and autocorrelation consistent coefficients

#Resumo da regressão com HAC
sum=summary(model2)
sum$coefficients = unclass(coeftest(model2, vcovHAC(model2, type = "
    HAC")))
```

## D.3 Código utilizado na RMLR

*Output* D.3: Input da RMLR

```
#Package
library(MASS)

#Regressão robusta com estimador M
model2_rlmM <-rlm(AMA_U_Proj12 ~ N_fr + cota + as.factor(n_ataque) +
    TMinF + TMaxQ + dias_pp,method ="M", data = DadosTotal)
summary(model2_rlmM)

#Regressão robusta com estimador MM
model2_rlmMM <-rlm(AMA_U_Proj12 ~ N_fr + cota + as.factor(n_ataque) +
    TMinF + TMaxQ + dias_pp,method ="MM", data = DadosTotal)
```

```
summary(model2_rlmMM)

#Função para  calcular o r-quadrado do modelo linear robusto
r2 <- function(x){
SSe <- sum((x$resid)^2);
observed <- x$resid+x$fitted;
SSt <- sum((observed-mean(observed))^2);
value <- 1-SSe/SSt;
return(value)}

r2_rlmM <- r2(model2_rlmM)
r2_rlmMM <- r2(model2_rlmMM)

#Função para  calcular o r-quadrado ajustado do modelo linear robusto
p<-(summary(model3)$df[1]) #número de coeficientes no modelo
n<-nrow(DadosTotal)
r2adj<-function(r2){
1-(1-r2)*((n-1)/(n-p-1))}

r2adj_rlmM<- r2adj(model2_rlmM)
r2adj_rlmMM<- r2adj(model2_rlmMM)
```